

CLOUD TECHNOLOGIES

A collection of technologies
relevant for clouds and big data

Gregor von Laszewski

Fugang Wang

Geoffrey C. Fox

laszewski@gmail.com

Cloud Technologies

Editor Gregor von Laszewski

(c) Gregor von Laszewski, 2018

Cloud Technologies

1 To Do

- [1.0.1 Bibtex Errors](#)
- [1.0.2 Check bibtex syntax](#)
- [1.0.3 Bibtex missing](#)
- [1.0.4 Revision requested](#)
- [1.0.5 Not Ready for Review](#)
- [1.0.6 Ready](#)

2 Preface

- [2.1 Notation](#)
- [2.2 Format](#)
- [2.3 Contributors](#)
- [2.4 Creating the Document](#)

3 Technologies

- [3.1 Accumulo](#)
- [3.2 ActiveBPEL](#)
- [3.3 ActiveMQ](#)
- [3.4 Aerobatic](#)
- [3.5 Agave](#)
- [3.6 Airavata](#)
- [3.7 AllegroGraph](#)
- [3.8 Amazon Dynamo](#)
- [3.9 Amazon Kinesis](#)
- [3.10 Amazon RDS](#) fa18-423-05
- [3.11 Amazon Redshift](#) fa18-423-06
- [3.12 Amazon Route 53](#)
- [3.13 Public Cloud: Amazon S3](#)
- [3.14 Public Cloud: Amazon SNS](#) fa18-523-52
- [3.15 Amazon](#)
- [3.16 Ambari](#) fa18-523-82
- [3.17 AMQP](#) fa18-523-70
 - [3.17.1 Advantages](#)
 - [3.17.2 Components:](#)
- [3.18 Ansible](#)
- [3.19 Any2Api](#)

3.20 Apache Ant	+
3.21 Apache Apex	+
3.22 Apache Arrow	+
3.23 Apache Beam	+
3.24 Apache Derby	+
3.25 Apache Flex	+
3.26 Apache HAWQ	+
3.27 Apache Knox	+
3.28 Apache OODT	+
3.29 Apache Ranger	+
3.30 Apache Tomcat	+
3.31 Apatar	+
3.32 appfog	+
3.33 AppScale	+
3.34 Askalon	+
3.35 Atmosphere	+
3.36 Avro	+
3.37 AWS OpsWorks	+
3.38 Azure Blob	+
3.39 Azure Data Factory	+
3.40 Azure Machine Learning	+
3.41 Azure Queues	+
3.42 Azure SQL	+
3.43 Azure Stream Analytics	+
3.44 Azure Table	+
3.45 Azure	+
3.46 Berkeley DB	+
3.47 Bioconductor	+
3.48 BioKepler	+
3.49 BitTorrent	+
3.50 Blaze	+
3.50.1 Old text	
3.50.2 New text	
3.51 Blazegraph	+
3.52 BlinkDB	+
3.53 Blueprints	+
3.54 Boto	+

3.55 Buildstep	⊕	
3.56 Caffe	⊕	
3.57 Cascading	fa18-523-65	⊕
3.58 Cassandra	⊕	
3.59 CDAP	⊕	
3.60 CDF	⊕	
3.61 CDMI	fa18-523-71	⊕
3.62 Celery	⊕	
3.63 Ceph	fa18-523-68	⊕
3.64 Chef	fa18-523-80	⊕
3.65 Cinder	⊕	
3.66 CINET	⊕	
3.67 Cisco Intelligent Automation for Cloud	⊕	
3.68 Cloud Foundry	fa18-523-64	⊕
3.69 Cloudability	fa18-523-86	⊕
3.70 CloudBees	⊕	
3.71 Cloudmesh	fa18-523-61	⊕
3.72 CloudML	⊕	
3.73 CloudStack	fa18-523-64	⊕
3.74 CNTK	⊕	
3.75 Cobbler	⊕	
3.76 CompLearn	⊕	
3.77 CoreOS	⊕	
3.78 Couchbase Server	fa18-523-85	⊕
3.79 CouchDB	fa18-423-03	⊕
3.79.1 Old		
3.79.2 New		
3.80 CUBRID	⊕	
3.81 CUDA	fa18-523-67	⊕
3.82 D3.js	⊕	
3.83 DAAL (Intel)	fa18-523-85	⊕
3.83.1 Old text		
3.83.2 New text		
3.84 Databus	⊕	
3.85 DataFu	fa18-523-61	⊕
3.86 Old		
3.87 New		

3.88 DataNucleus	Fa18-523-73	⊕
3.89 DataTurbine	fa18-423-02	⊕
3.90 DB2		⊕
3.91 DC.js	fa18-523-58	⊕
3.92 DevOpSlang		⊕
3.93 Disco	fa18-523-63	⊕
	3.93.1 Old Text	
	3.93.2 New Text	
3.94 Distributed Coordination		⊕
3.95 DL4j		⊕
3.96 Docker Compose	fa18-523-60	⊕
3.97 Docker (Machine, Swarm)		⊕
3.98 Dokku	fa18-523-57	⊕
3.99 dotCloud		⊕
3.100 Dream:Lab	fa18-523-79	⊕
3.101 Drill	fa18-523-81	⊕
3.102 Dryad	fa18-523-58	⊕
3.103 e-Science Central	fa18-523-68	⊕
3.104 EclipseLink		⊕
3.105 Eduroam		⊕
3.106 Ehcache		⊕
3.107 Elasticsearch	fa18-523-70	⊕
3.108 Engine Yard		⊕
3.109 Espresso	fa18-523-79	⊕
3.110 Eucalyptus		⊕
3.111 Event Hubs	fa18-523-57	⊕
3.112 f4		⊕
3.113 Facebook Corona		⊕
3.114 Facebook Puma/Ptail/Scribe/ODS		⊕
3.115 Facebook Tao	fa18-523-86	⊕
3.116 Facebook Tupperware		⊕
3.117 Fiddler		⊕
3.118 FITS		⊕
3.119 Flink Streaming	fa18-523-80	⊕
3.120 Floe		⊕
3.121 Flume		⊕
3.122 Foreman		⊕

3.123 FTP	fa18-523-63	⊕
3.123.1 Old Text		
3.123.2 New Text		
3.124 FUSE		⊕
3.125 Galaxy		⊕
3.126 Galera Cluster		⊕
3.127 Galois		⊕
3.128 Ganglia		⊕
3.129 Genesis		⊕
3.130 GFFS	fa18-523-86	⊕
3.131 Giraffe		⊕
3.132 Giraph	fa18-523-64	⊕
3.133 Gitreceive		⊕
3.134 Globus Online - GridFTP	Fa18-523-74	⊕
3.134.1 NEW TEXT		
3.135 Globus Tools		⊕
3.136 Gluster		⊕
3.137 Google and other public Clouds		⊕
3.138 Google App Engine		⊕
3.139 Google BigQuery	fa18-523-63	⊕
3.139.1 Old Text		
3.139.2 New Text		
3.140 Google Bigtable :Smiley:	fa18-423-06	⊕
3.141 Google Chubby	FA18-523-53	⊕
3.142 Google Cloud Dataflow		⊕
3.143 Google Cloud Dataflow		
3.143.1 Duplicated entry: merge		
3.144 Networking: Google Cloud DNS		⊕
3.145 Google Cloud Machine Learning		⊕
3.146 Google Cloud SQL	fa18-523-58	⊕
3.147 Google Cloud Storage		⊕
3.148 Google DataStore		⊕
3.149 Google Dremel		⊕
3.150 Google F1		⊕
3.151 Google FlumeJava		⊕
3.152 Google Fusion Tables	fa18-523-71	⊕
3.153 Google Kubernetes	fa18-523-56	⊕

3.154 Google MillWheel	⊕
3.155 Google Omega	⊕
3.156 Google Prediction API & Translation API	⊕
3.157 Google Pub Sub	⊕
3.158 Gora (general object from NoSQL)	⊕
3.159 Granules	⊕
3.160 GraphBuilder (Intel)	⊕
3.161 GraphChi	⊕
3.162 graphdb	⊕
3.163 GraphLab	fa18-423-05
3.164 GraphX	⊕
3.165 Graylog	⊕
3.166 H-Store	fa18-523-62
3.167 H2O	⊕
3.168 Hadoop	⊕
3.169 HadoopDB	⊕
3.170 Hama	⊕
3.171 Harp	⊕
3.172 Haystack	⊕
3.173 Hazelcast	⊕
3.174 HBase	fa18-423-02
3.175 HCatalog	fa18-523-81
3.176 HDF	fa18-523-69
3.177 HDFS	⊕
3.178 Helix	fa18-523-62
3.179 Heroku	fa18-523-67
3.180 Hibernate	⊕
3.181 Hive	fa18-423-05
3.182 (a) publish-subscribe: MPI	⊕
3.183 HPX-5	⊕
3.184 HTCondor	⊕
3.185 HTTP	⊕
3.186 HUBzero	⊕
3.187 Hyper-V	fa18-523-81
3.187.1 Hyper-V Architecture:	⊕
3.187.2 Advantages of Hyper-V over other hypervisors:	⊕
3.188 IBM BlueMix	fa18-423-06

3.189 IBM Cloudant	⊕
3.190 IBM dashDB	⊕
3.191 IBM Spectrum Scale, formerly GPFS	⊕
3.192 IBM System G	⊕
3.192.1 Old: IBM Watson	
3.192.2 New: IBM Watson	
3.193 ImageJ	⊕
3.194 Impala	⊕
3.195 Inca	⊕
3.196 InCommon	⊕
3.197 Infinispan	⊕
3.198 iRODS	⊕
3.199 JClouds	fa18-523-68
3.200 Jelastic	⊕
3.201 Jena	fa18-523-56
3.202 JGroups	⊕
3.203 Jitterbit	⊕
3.204 JMS	⊕
3.205 Juju	fa18-523-83
3.206 Jupyter and IPython	⊕
3.207 Kafka	fa18-523-65
3.208 Karajan	⊕
3.209 Kepler	⊕
3.210 Kestrel	⊕
3.211 KeystoneML	⊕
3.212 Kibana	⊕
3.213 Kite	⊕
3.214 KVM	⊕
3.215 Kyoto Cabinet	⊕
3.216 Kyoto/Tokyo Cabinet	⊕
3.217 Lambda	⊕
3.218 LDAP	⊕
3.219 LevelDB	⊕
3.220 Libcloud	fa18-523-59
3.221 Libvirt	⊕
3.222 Ligra	⊕
3.223 LinkedIn	⊕

3.224 Linux-Vserver	fa18-523-60	⊕
3.225 Llama		⊕
3.226 LMDB (key value)	fa18-423-06	⊕
3.227 Logstash		⊕
3.228 Lucene	fa18-523-79	⊕
3.229 Lumberyard		⊕
3.230 Lustre	fa18-523-82	⊕
3.231 LXC	fa18-523-85	⊕
3.231.1 Old text		
3.231.2 New text		
3.232 LXD	fa18-523-71	⊕
3.233 Mahout		⊕
3.234 MapGraph		⊕
3.235 Marionette Collective		⊕
3.236 Medusa-GPU		⊕
3.237 Megastore and Spanner		⊕
3.238 Memcached	fa18-523-52	⊕
3.239 Mesos	fa18-523-79	⊕
3.240 MLbase	fa18-523-84	⊕
3.241 MLLib		⊕
3.242 mipy	fa18-523-68	⊕
3.243 Moab		⊕
3.244 MongoDB		⊕
3.245 MQTT		⊕
3.246 MR-MPI		⊕
3.247 MRQL	fa18-523-69	⊕
3.248 MySQL	fa18-523-60	⊕
3.249 N1QL		⊕
3.250 Nagios		⊕
3.251 Naiad		⊕
3.252 NaradaBrokering		⊕
3.253 Neo4j		⊕
3.254 Neptune	fa18-523-73	⊕
3.255 NetCDF		⊕
3.256 Netty		⊕
3.257 NiFi (NSA)	fa18-523-56	⊕
3.258 Nimbus	fa18-523-65	⊕

3.259 Ninefold		+
3.260 NWB	fa18-523-69	+
3.261 OCCI	fa18-523-80	+
3.262 ODBC/JDBC		+
3.263 ODE		+
3.264 Omid		+
3.265 OODT		+
3.266 Oozie		+
3.267 Open MPI		+
3.268 OpenCV	fa18-423-02	+
3.269 OPeNDAP	fa18-523-72	+
3.270 OpenID	fa18-523-63	+
3.270.1 Old Text		
3.270.2 New Text		
3.271 OpenJPA	Fa18-523-74	+
3.271.1 NEW TEXT		
3.272 OpenNebula	fa18-523-68	+
3.273 OpenPBS		+
3.274 OpenRefine		+
3.275 OpenStack Heat	fa18-523-58	+
3.276 OpenStack Ironic		+
3.277 OpenStack Keystone	fa18-523-59	+
3.278 OpenStack		+
3.279 OpenTOSCA		+
3.280 OpenVZ	fa18-523-71	+
3.281 Oracle PGX		+
3.282 Oracle	fa18-423-06	+
3.283 ORC		+
3.284 OSGi	fa18-523-85	+
3.285 Parasol		+
3.286 Parquet		+
3.287 pbdr	FA18-523-53	+
3.288 Pegasus		+
3.289 Pentaho		+
3.290 PetSc	fa18-523-82	+
3.291 Phoenix	fa18-523-72	+
3.292 Pig		+

3.293 Pilot Jobs	⊕
3.294 Pivotal Gemfire	⊕
3.295 Pivotal GPOLOAD/GPFDIST	⊕
3.296 Pivotal Greenplum	⊕
3.297 Pivotal	⊕
3.298 PLASMA MAGMA	fa18-523-60
3.299 point-to-point	⊕
3.300 PolyBase	fa18-523-69
3.301 PostgreSQL	⊕
3.301.1 Old Entry	⊕
3.301.2 New entry	⊕
3.302 Potree	⊕
3.303 Pregel	fa18-523-82
3.304 Presto	⊕
3.305 Protobuf	fa18-523-56
3.306 (b) publish-subscribe: Big Data	⊕
3.307 Puppet	fa18-523-82
3.308 PyBrain	fa18-523-59
3.309 QEMU	⊕
3.310 QPid	⊕
3.311 R	fa18-523-66
3.312 RabbitMQ	fa18-523-74
3.313 Rasdaman	fa18-523-70
3.313.1 Key Features	⊕
3.314 Razor	⊕
3.315 RCFile	⊕
3.316 Red Hat OpenShift	⊕
3.317 Redis	Fa18-523-74
3.317.1 NEW TEXT	⊕
3.318 Reef	⊕
3.319 Riak	fa18-523-57
3.320 rkt	fa18-523-64
3.321 Robot Operating System (ROS)	⊕
3.322 Rocks	⊕
3.323 RYA	⊕
3.324 S4	⊕
3.325 Saga	⊕

3.326 Sahara	+
3.327 SaltStack	fa18-523-86
3.328 SAML OAuth	+
3.329 Samza	+
3.330 SAP HANA	fa18-523-83
3.331 Sawzall	+
3.332 Scalapack	+
3.333 Scalding	+
3.334 SciDB	+
3.335 scikit-learn	+
3.336 Security and Privacy	+
3.337 Sentry	fa18-523-65
3.338 Sesame	+
3.339 SGE - Univa Grid Engine	fa18-523-83
3.340 Shark	fa18-523-72
3.341 Slurm	fa18-523-83
3.342 Snort	fa18-523-66
3.343 Solandra	+
3.344 Solr	F18-523-53
3.345 Spark SQL	+
3.346 Spark Streaming	fa18-523-67
3.347 Spark	+
3.348 Splunk	+
3.349 SQL Server	fa18-523-57
3.350 SQLite	fa18-523-61
3.350.1 Previous text	+
3.350.2 New	+
3.351 Sqoop	+
3.352 Sqrrl	+
3.353 SSH	+
3.354 Stackato	+
3.355 Stomp	+
3.356 Storm	+
3.357 Apache Flink	+
3.358 Summingbird	+
3.359 Swift	+
3.360 Tableau	+

3.361 Tajo	fa18-523-80	⊕
3.362 Talend		⊕
3.363 Taverna	fa18-523-66	⊕
3.364 TensorFlow	fa18-423-02	⊕
3.365 Terraform	fa18-523-62	⊕
3.366 Tez		⊕
3.367 Theano		⊕
3.368 three.js		⊕
3.369 Thrift	fa18-423-08	⊕
3.370 Tika		⊕
3.371 TinkerPop		⊕
3.372 Titan:db	fa18-523-52	⊕
3.373 Torch		⊕
3.374 Torque		⊕
3.375 TOSCA		⊕
3.376 Totem		⊕
3.377 Triana		⊕
3.378 Trident		⊕
3.379 Twister	fa18-523-67	⊕
3.380 Twitter Heron	fa18-423-05	⊕
3.381 Tycoon	fa18-523-52	⊕
3.382 Tyrant	fa18-523-61	⊕
3.383 old text		
3.384 New text		
3.385 Ubuntu MaaS	fa18-523-84	⊕
3.386 UIMA	fa18-523-62	⊕
3.387 VirtualBox	fa18-423-03	⊕
3.388 VMware ESXi		⊕
3.389 Voldemort	fa18-523-70	⊕
	3.389.1 Disadvantages of Voldemort:	
3.390 VoltDB		⊕
3.391 vSphere and vCloud	fa18-523-85	⊕
3.392 Whirr		⊕
3.393 Winery	fa18-523-79	⊕
3.394 Wink	fa18-523-84	⊕
3.395 Xcat		⊕
3.396 Xen		⊕

3.397 Yarcdata	⊕
3.398 Yarn	⊕
3.399 Zeppelin	⊕
3.400 ZeroMQ	fa18-523-79
3.401 ZHT	⊕
3.402 Zookeeper	⊕
4 Incomming	
4.1 AlibabaCloud	⊕
4.2 Alluxio	⊕
4.3 AWS API Gateway	⊕
4.4 Amazon Aurora	⊕
4.5 Amazon CloudFront	⊕
4.6 AWS CodeStar	⊕
4.7 AWS DeepLens	⊕
4.8 Amazon DynamoDB	⊕
4.9 Amazon EC2	⊕
4.10 Amazon Elastic Beanstalk	⊕
4.11 Amazon Fargate	⊕
4.12 Amazon Glacier	⊕
4.13 Amazon Lightsail	⊕
4.14 Amazon Machine Learning	⊕
4.15 Amazon RDS	⊕
4.16 Amazon Redshift	⊕
4.17 Amazon S3	⊕
4.18 Amazon VPC	⊕
4.19 Ansible	⊕
4.20 Apache Accumulo	⊕
4.21 Apache Ambari	⊕
4.22 Apache Atlas	⊕
4.23 Apache Avro	⊕
4.24 Apache Chukwa	⊕
4.25 Apache CloudStack	⊕
4.26 Apache Couch DB	⊕
4.27 Apache Curator	⊕
4.28 Apache Delta Cloud	⊕
4.29 Apache Drill	⊕
4.30 Apache Geode	⊕

4.31 Apache Ignite	+
4.32 Apache Impala	+
4.33 Apache Karaf	+
4.34 Apache Kylin	+
4.35 Apache Mahout	+
4.36 Apache Mesos	+
4.37 Apache Phoenix	+
4.38 Apache Whirr	+
4.39 Apache Zookeeper	+
4.40 Apatar	+
4.41 AppFog	+
4.42 Appscale	+
4.43 Apttus	+
4.44 ArangoDB	+
4.45 Amazon Athena	+
4.46 AtomSphere	+
4.47 Azure	+
4.48 Azure Blob Storage	+
4.49 Azure Cosmos DB	+
4.50 Backblaze	+
4.51 BigML	+
4.52 Apache BigTop	+
4.53 Blockchain	+
4.54 IBM BlueMix	+
4.55 BMC Multi-Cloud	+
4.56 Caffe	+
4.57 Apache Carbondata	+
4.58 Cascading	+
4.59 CensOS Project	+
4.60 Clive	+
4.61 Clojure	+
4.62 Cloud AutoML	+
4.63 CloudHub	+
4.64 Cloudlet	+
4.65 CloudTrail	+
4.66 CloudWatch	+
4.67 Microsoft Cognitive Toolkit	+

4.68 Cognito	+
4.69 ConnectTheDots	+
4.70 CouchDB	+
4.71 Databricks	+
4.72 Datalab	+
4.73 Datameer	+
4.74 DBI	+
4.75 DBplyr	+
4.76 Data Virtualization	+
4.77 Distributed Machine Learning Tool Kit	+
4.78 docker	+
4.79 Dokku	+
4.80 Drake	+
4.81 Google Dremel	+
4.82 Druid	+
4.83 IBM Data Science Experience	+
4.84 Edge Computing	+
4.85 Apache Edgent	+
4.86 Elasticsearch	+
4.87 ELK Stack	+
4.88 Amazon EMR	+
4.89 ESRI	+
4.90 Ethereum	+
4.91 Firebase	+
4.92 Firepad	+
4.93 Fission	+
4.94 Fluentd	+
4.95 FoundationBenchmarks	+
4.96 Future Grid	+
4.97 Google Cloud Platform - Big data solutions	+
4.98 Google Cloud Platform - Cloud Dataproc	+
4.99 Google Genomics	+
4.100 Gephi	+
4.101 GitHub Developer	+
4.102 Apache Gobblin	+
4.103 Google App Engine	+
4.104 Google BigQuery	+

4.105 Cloud Bigtable	+
4.106 Google Compute Engine	+
4.107 Google Docs	+
4.108 Google Firebase	+
4.109 Google Load Balancing	+
4.110 Google Stackdriver	+
4.111 Apache Gossip	+
4.112 GraphQL	+
4.113 AWS Greengrass	+
4.114 H2O	+
4.115 Apache Hadoop	+
4.116 HBase	+
4.117 HCatalog	+
4.118 HPCC Systems	+
4.119 Hue	+
4.120 Hyperledger Burrow	+
4.121 Hyperledger Fabric	+
4.122 Hyperledger Indy	+
4.123 Hyperledger Iroha	+
4.124 Hyperledger Sawtooth	+
4.125 IBM Big Replicate	+
4.126 IBM Cloud	+
4.127 IBM Db2 Big Sql	+
4.128 ID2020	+
4.129 IICS	+
4.130 Instabug	+
4.131 Intel Cloud Finder	+
4.132 Jaspersoft	+
4.133 JavaScript	+
4.134 Jelastic	+
4.135 JMP	+
4.136 Kafka	+
4.137 Keras	+
4.138 KNIME	+
4.139 Kubernetes	+
4.140 Kudu	+
4.141 Kylin	+

4.142 LightGBM	+
4.143 Lingual	+
4.144 LinkedIn WhereHows	+
4.145 Linode	+
4.146 Logicalglue	+
4.147 Lumify	+
4.148 Apache Mahout	+
4.149 MapBox	+
4.150 MariaDB	+
4.151 Mesosphere	+
4.152 Metron	+
4.153 Apache Milagro	+
4.154 mLab	+
4.155 MonetDB	+
4.156 MongoDB	+
4.157 Morpheus	+
4.158 Microsoft Visual Studio	+
4.159 neo4j	+
4.160 Neptune	+
4.161 Netflix	+
4.162 Apache NiFi	+
4.163 Node.js	+
4.164 ODBC	+
4.165 OneDrive	+
4.166 Oozie	+
4.167 Openchain	+
4.168 OpenDaylight	+
4.169 OpenNebula	+
4.170 OpenNN	+
4.171 Open Refine	+
4.172 openVZ	+
4.173 Oracle Big Data Cloud Service	+
4.174 Oracle Coherence - DataGrid	+
4.175 Oracle Nosql Database	+
4.176 Orange	+
4.177 OrientDB hid-SP18-520	+
4.178 Owncloud	+

4.179 Paxata	+
4.180 Pig	+
4.181 Pivotal	+
4.182 Pivotal Rabbit MQ	+
4.183 Pool	+
4.184 Apache PredictionIO	+
4.185 Presto	+
4.186 PubNub	+
4.187 Pulsar	+
4.188 Puppet	+
4.189 PyTorch	+
4.190 Qubole Data Service	+
4.191 RabbitMQ	+
4.192 Rackspace	+
4.193 Ranger	+
4.194 RapidMiner	+
4.195 Redis	+
4.196 RightScale Cloud Management	+
4.197 Ripple Transaction Protocol	+
4.198 Amazon SageMaker	+
4.199 Sales Cloud	+
4.200 Apache Samoa	+
4.201 Scikit-learn	+
4.202 Scribe	+
4.203 SETI @ Home	+
4.204 ShareLatex	+
4.205 Share Point	+
4.206 Skytap	+
4.207 Apache Solr	+
4.208 SpagoBI	+
4.209 Google Cloud Spanner	+
4.210 Spinnaker	+
4.211 SQLite	+
4.212 Swoop	+
4.213 Stardog	+
4.214 Synthea	+
4.215 Synthetic Data Vault	+

4.216 Apache SystemML	+
4.217 Tableau	+
4.218 Talend	+
4.219 TensorFlow	+
4.220 Teradata Intelliflex	+
4.221 Teradata Intellibase	+
4.222 Teradata Kylo	+
4.223 Theano	+
4.224 The GO Programming Language	+
4.225 Tibco DataSynapse GridServer	+
4.226 TokuDB	+
4.227 TreasureData	+
4.228 Twilio	+
4.229 US Consumer Financial Protection Bureau	+
4.230 Google Vision	+
4.231 Weka	+
4.232 The World Bank	+
4.233 WSO2 Analytics	+
4.234 XGBoost	+
4.235 Zepplin	+
4.236 Zmanda	+
Refernces	+

1 To Do



Todos are integrated in the text marked with TODOs. some of them have a warning icon. Others will be listed here.

Open Pull Requests

Link	Check	Title	user
649	check	made the changes we talked about today plasma-magma, mysql	IzoldalU
640	check	Update references-fa18-523-62.bib	manekbahl
639	check	Update h-store.md	manekbahl
638	check	fa18-523-68 Tech Summaries - Requested Fixes	sakkas20
629	check	Create references-fa18-423-05.bib	yixihu
616	check	Update gffs.md	jeffliuhai
		Update	

614	check	references-fa18-523-86.bib	jeffliuhai
589	check	Update osgi.md	Hidgons
587	check	Update tycoon.md	avheine
531	check	Update hive.md	yixihu
525	check	Update graphlab.md	yixihu
515	check	Update bittorrent.md	wangton

1.0.1 Bibtex Errors

- fa18-423-05-amazon not found
- fa18-423-06-Redshift-Review not found
- fa18-423-06-Redshift-Review not found
- fa18-423-06-Redshift-Review not found
- fa18-423-03-berkeleydb not found
- www-couchdb.apache not found
- www-couchdb.apache not found
- fa18-523-68-www-escience-central-about not found
- fa18-523-86-www not found
- fa18-523-86-www not found
- www-rdc114 not found
- www-ftp-man not found
- fa18-523-86-www not found
- fa18-523-86-www-GFFS not found
- fa18-523-86-www-GFFS not found
- fa18-523-86-www-GFFS-Wiki not found
- Inside_look_at_Google_Bigquery not found
- fa18-423-05-turi not found

- fa18-423-05-techcrunch not found
- fa18-423-05-www-oreilly not found
- fa18-423-02-hbase-org not found
- fa18-423-02-hbase not found
- fa18-423-05-hive-apache not found
- fa18-423-05-intellipaat not found
- fa18-423-05-intellipaat not found
- fa18-423-05-whizlabs not found
- fa18-423-05-whizlabs not found
- fa18-423-05-whizlabs not found
- fa18-423-05-intellipaat not found
- fa18-423-05-intellipaat not found
- fa18-523-68-lsmael-jclouds not found
- fa18-523-56-www-Jena-ARQquery not found
- fa18-423-06-www-LMBD not found
- Harkness-2017 not found
- www-Oracle-DB not found
- www-Oracle-Defintion not found
- www-Oracle-Amazon not found
- www-Oracle-Amazon not found
- www-Oracle-Autonomous not found
- www-Oracle-Autonomous not found
- fa18-523-www-protobuf-googleinterchangeformatProtocol not found
- fa18-523-www-protobuf-googleinterchangeformatProtocol not found
- fa18-523-86-www not found
- fa18-523-86-www not found
- fa18-523-86-www not found
- fa18-523-52-architectural not found
- fa18-423-05-apache-github not found
- fa18-423-03-virtualbox not found

1.0.2 Check bibtex syntax

1.0.2.1 bib/references-fa18-423-02.bib

- Name “Tony Fountain, Sameer Tilak, Peter Shin, Michael Nekrasov” has too many commas: skipping name

1.0.2.2 bib/references-fa18-523-66.bib

- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-r-wiki’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-taverna-intro’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-derby-wiki’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-r-project’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-snort’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-apache-derby’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-taverna-wiki’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-taverna-pred-analytics’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-ms-sql-server’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-apache-taverna’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-snort-wiki’
- Overwriting field ‘month’ with month value from field ‘date’ for entry ‘www-oracle-database’

1.0.2.3 bib/references.bib

- Overwriting field ‘year’ with year value from field ‘date’ for entry ‘book-mesos-Ignazio-2016’
- Overwriting field ‘year’ with year value from field ‘date’ for entry ‘hid-sp18-404-BlogsApache2014’

- Overwriting field 'year' with year value from field 'date' for entry 'paper-azure-1'
- Overwriting field 'year' with year value from field 'date' for entry 'paper-samza-1'
- ISBN '9351105164 and 9789351105169' in entry 'chef-book' is invalid - run biber with '--validate_datamodel' for details.
- Datamodel: Entry 'amazon-elastic-beanstalk-book' (bib/references.bib): Invalid format '2011-8-06' of date field 'date' - ignoring
- Overwriting field 'year' with year value from field 'date' for entry 'www-jelastic-2'
- Overwriting field 'month' with month value from field 'date' for entry 'www-jelastic-2'
- Overwriting field 'year' with year value from field 'date' for entry 'www-kibana-wiki'
- Overwriting field 'month' with month value from field 'date' for entry 'www-kibana-wiki'
- Overwriting field 'year' with year value from field 'date' for entry 'wrong-label-2017i'
- Overwriting field 'month' with month value from field 'date' for entry 'wrong-label-2017i'
- Overwriting field 'year' with year value from field 'date' for entry 'www-samza-3'
- Overwriting field 'month' with month value from field 'date' for entry 'www-samza-3'
- Overwriting field 'year' with year value from field 'date' for entry 'netty-book'
- ISBN '013313573X, 9780133135732' in entry 'nagios-book' is invalid - run biber with '--validate_datamodel' for details.
- ISBN '978-4919-6573-3' in entry 'www-cloudfoundry-book' is invalid - run biber with '--validate_datamodel' for details.
- Overwriting field 'year' with year value from field 'date' for entry 'paper-openvz-1'
- Overwriting field 'year' with year value from field 'date' for entry 'www-jupyter-wiki'
- Overwriting field 'month' with month value from field 'date' for entry 'www-jupyter-wiki'

- Overwriting field 'year' with year value from field 'date' for entry 'www-samza-4'
- Overwriting field 'month' with month value from field 'date' for entry 'www-samza-4'
- ISBN '1933988940, 9781933988948' in entry 'ActiveMQ-book' is invalid - run biber with '--validate_datamodel' for details.
- Overwriting field 'year' with year value from field 'date' for entry 'www-kibana-4'
- Overwriting field 'month' with month value from field 'date' for entry 'www-kibana-4'
- Overwriting field 'year' with year value from field 'date' for entry 'book-scikit-learn'
- Overwriting field 'year' with year value from field 'date' for entry 'www-azure-3'
- Overwriting field 'month' with month value from field 'date' for entry 'www-azure-3'
- Overwriting field 'year' with year value from field 'date' for entry 'book-Cinder'
- ISBN '978-1-4799-8430-55' in entry 'ligra-paper-2' is invalid - run biber with '--validate_datamodel' for details.
- ISBN '978-1-78439-460-8, 1784394602' in entry 'SaltStack-book' is invalid - run biber with '--validate_datamodel' for details.
- Overwriting field 'year' with year value from field 'date' for entry 'book-greenplum-gollapudi2013'
- Overwriting field 'month' with month value from field 'date' for entry 'book-greenplum-gollapudi2013'
- Overwriting field 'year' with year value from field 'date' for entry 'paper-thrift'
- ISBN '331913700X, 9783319137001' in entry 'cloud-portability-book' is invalid - run biber with '--validate_datamodel' for details.
- BibTeX subsystem: warning: comma(s) at end of name (removing)
- BibTeX subsystem: author, warning: comma(s) at end of name (removing)

1.0.3 Bibtex missing

- chapters/tech/amazon-rds.md
- chapters/tech/amazon-redshift.md
- chapters/tech/aws-elastic-beanstalk.md
- chapters/tech/berkeley-db.md
- chapters/tech/e-science-central.md
- chapters/tech/ftp.md
- chapters/tech/gffs.md
- chapters/tech/graphlab.md
- chapters/tech/hbase.md
- chapters/tech/hive.md
- chapters/tech/jclouds.md
- chapters/tech/jena.md
- chapters/tech/lmdb-key-value.md
- chapters/tech/neptune.md
- chapters/tech/oracle.md
- chapters/tech/protobuf.md
- chapters/tech/titan-db.md
- chapters/tech/twitter-heron.md
- chapters/tech/virtualbox.md

1.0.4 Revision requested

- tech/ambari.md: Ambari fa18-523-82
- tech/cDMI.md: CDMI fa18-523-71
- tech/gffs.md: GFFS fa18-523-86
- tech/ibm-system-g.md: IBM System G
- tech/lxd.md: LXD fa18-523-71
- tech/openvz.md: OpenVZ fa18-523-71
- tech/rabbitmq.md: RabbitMQ fa18-523-74
- tech/saltstack.md: SaltStack fa18-523-86

1.0.5 Not Ready for Review

- incomming/dremel.md: Google Dremel
- incomming/genomics.md: Google Genomics

- incomming/h2o.md: H2O
- tech/bioconductor.md: Bioconductor fa18-xxx-xx
- tech/cassandra.md: Cassandra
- tech/couchbase-server.md: Couchbase Server fa18-523-85
- tech/crunch.md: Crunch f18-523-53
- tech/google-cloud-dataflow.md: Google Cloud Dataflow
- tech/mlpy.md: mlpy fa18-523-68
- tech/pbdr.md: pbdR FA18-523-53
- tech/solr.md: Solr F18-523-53
- tech/vsphere-and-vcloud.md: vSphere and vCloud fa18-523-85

1.0.6 Ready

- incomming/spanner.md: Google Cloud Spanner
- incomming/vision.md: Google Vision
- tech/amazon-rds.md: Amazon RDS fa18-423-05
- tech/amazon-redshift.md: Amazon Redshift fa18-423-06
- tech/amazon-sns.md: Public Cloud: Amazon SNS fa18-523-52
- tech/ambari.md: Ambari fa18-523-82
- tech/amqp.md: AMQP fa18-523-70
- tech/apache-apex.md: Apache Apex fa18-523-58
- tech/apache-arrow.md: Apache Arrow fa18-423-08
- tech/apache-beam.md: Apache Beam fa18-523-81
- tech/apache-derby.md: Apache Derby fa18-523-66
- tech/apache-hawq.md: Apache HAWQ fa18-523-84
- tech/apache-oodt.md: Apache OODT fa18-523-81
- tech/appfog.md: appfog fa18-523-72
- tech/atmosphere.md: Atmosphere fa18-523-58
- tech/aws-elastic-beanstalk.md: AWS Elastic Beanstalk fa18-423-03
- tech/azure-data-factory.md: Azure Data Factory fa18-523-84
- tech/azure-sql.md: Azure SQL fa18-523-59
- tech/azure-table.md: Azure Table fa18-523-73
- tech/berkeley-db.md: Berkeley DB fa18-423-03

- tech/bittorrent.md: BitTorrent Fa18-523-73
- tech/blaze.md: Blaze fa18-523-85
- tech/blinkdb.md: BlinkDB fa18-523-82
- tech/cascading.md: Cascading fa18-523-65
- tech/cdmi.md: CDMI fa18-523-71
- tech/ceph.md: Ceph fa18-523-68
- tech/chef.md: Chef fa18-523-80
- tech/cloud-foundry.md: Cloud Foundry fa18-523-64
- tech/cloudability.md: Cloudability fa18-523-86
- tech/cloudmesh.md: Cloudmesh fa18-523-61
- tech/cloudstack.md: CloudStack fa18-523-64
- tech/couchdb.md: CouchDB fa18-423-03
- tech/cuda.md: CUDA fa18-523-67
- tech/daal-intel.md: DAAL (Intel) fa18-523-85
- tech/datafu.md: DataFu fa18-523-61
- tech/datanucleus.md: DataNucleus Fa18-523-73
- tech/dataturbine.md: DataTurbine fa18-423-02
- tech/dc.js.md: DC.js fa18-523-58
- tech/disco.md: Disco fa18-523-63
- tech/docker-compose.md: Docker Compose fa18-523-60
- tech/dokku.md: Dokku fa18-523-57
- tech/dream-lab.md: Dream:Lab fa18-523-79
- tech/drill.md: Drill fa18-523-81
- tech/dryad.md: Dryad fa18-523-58
- tech/e-science-central.md: e-Science Central fa18-523-68
- tech/elasticsearch.md: Elasticsearch fa18-523-70
- tech/espresso.md: Espresso fa18-523-79
- tech/event-hubs.md: Event Hubs fa18-523-57
- tech/facebook-tao.md: Facebook Tao fa18-523-86
- tech/flink-streaming.md: Flink Streaming fa18-523-80
- tech/ftp.md: FTP fa18-523-63
- tech/giraph.md: Giraph fa18-523-64
- tech/globus-online-gridftp.md: Globus Online - GridFTP Fa18-523-74
- tech/google-bigquery.md: Google BigQuery fa18-523-63

- tech/google-chubby.md: Google Chubby FA18-523-53
- tech/google-cloud-sql.md: Google Cloud SQL fa18-523-58
- tech/google-fusion-tables.md: Google Fusion Tables fa18-523-71
- tech/google-kubernetes.md: Google Kubernetes fa18-523-56
- tech/graphlab.md: GraphLab fa18-423-05
- tech/h-store.md: H-Store fa18-523-62
- tech/hbase.md: HBase fa18-423-02
- tech/hcatalog.md: HCatalog fa18-523-81
- tech/hdf.md: HDF fa18-523-69
- tech/helix.md: Helix fa18-523-62
- tech/heroku.md: Heroku fa18-523-67
- tech/hive.md: Hive fa18-423-05
- tech/hyper-v.md: Hyper-V fa18-523-81
- tech/ibm-bluemix.md: IBM BlueMix fa18-423-06
- tech/jclouds.md: JClouds fa18-523-68
- tech/jena.md: Jena fa18-523-56
- tech/juju.md: Juju fa18-523-83
- tech/kafka.md: Kafka fa18-523-65
- tech/libcloud.md: Libcloud fa18-523-59
- tech/linux-vserver.md: Linux-Vserver fa18-523-60
- tech/lmdb-key-value.md: LMDB (key value) fa18-423-06
- tech/lucene.md: Lucene fa18-523-79
- tech/lustre.md: Lustre fa18-523-82
- tech/lxc.md: LXC fa18-523-85
- tech/lxd.md: LXD fa18-523-71
- tech/memcached.md: Memcached fa18-523-52
- tech/mesos.md: Mesos fa18-523-79
- tech/mlbase.md: MLbase fa18-523-84
- tech/mrql.md: MRQL fa18-523-69
- tech/mysql.md: MySQL fa18-523-60
- tech/neptune.md: Neptune fa18-523-73
- tech/nifi-nsa.md: NiFi (NSA) fa18-523-56
- tech/nimbus.md: Nimbus fa18-523-65
- tech/nwb.md: NWB fa18-523-69

- tech/occi.md: OCCI fa18-523-80
- tech/opencv.md: OpenCV fa18-423-02
- tech/opendap.md: OPeNDAP fa18-523-72
- tech/openid.md: OpenID fa18-523-63
- tech/openjpa.md: OpenJPA Fa18-523-74
- tech/opennebula.md: OpenNebula fa18-523-68
- tech/openstack-heat.md: OpenStack Heat fa18-523-58
- tech/openstack-keystone.md: OpenStack Keystone fa18-523-59
- tech/openvz.md: OpenVZ fa18-523-71
- tech/oracle.md: Oracle fa18-423-06
- tech/osgi.md: OSGi fa18-523-85
- tech/petsc.md: PetSc fa18-523-82
- tech/phoenix.md: Phoenix fa18-523-72
- tech/plasma-magma.md: PLASMA MAGMA fa18-523-60
- tech/polybase.md: PolyBase fa18-523-69
- tech/pregel.md: Pregel fa18-523-82
- tech/protobuf.md: Protobuf fa18-523-56
- tech/puppet.md: Puppet fa18-523-82
- tech/pybrain.md: PyBrain fa18-523-59
- tech/r.md: R fa18-523-66
- tech/rasdaman.md: Rasdaman fa18-523-70
- tech/redis.md: Redis Fa18-523-74
- tech/riak.md: Riak fa18-523-57
- tech/rkt.md: rkt fa18-523-64
- tech/sap-hana.md: SAP HANA fa18-523-83
- tech/sentry.md: Sentry fa18-523-65
- tech/sge.md: SGE - Univa Grid Engine fa18-523-83
- tech/shark.md: Shark fa18-523-72
- tech/slurm.md: Slurm fa18-523-83
- tech/snort.md: Snort fa18-523-66
- tech/spark-streaming.md: Spark Streaming fa18-523-67
- tech/sql-server.md: SQL Server fa18-523-57
- tech/sqlite.md: SQLite fa18-523-61
- tech/tajo.md: Tajo fa18-523-80
- tech/taverna.md: Taverna fa18-523-66
- tech/tensorflow.md: TensorFlow fa18-423-02

- tech/terraform.md: Terraform fa18-523-62
- tech/thrift.md: Thrift fa18-423-08
- tech/titan-db.md: Titan:db fa18-523-52
- tech/twister.md: Twister fa18-523-67
- tech/twitter-heron.md: Twitter Heron fa18-423-05
- tech/tycoon.md: Tycoon fa18-523-52
- tech/tyrant.md: Tyrant fa18-523-61
- tech/ubuntu-maas.md: Ubuntu MaaS fa18-523-84
- tech/uima.md: UIMA fa18-523-62
- tech/virtualbox.md: VirtualBox fa18-423-03
- tech/voldemort.md: Voldemort fa18-523-70
- tech/winery.md: Winery fa18-523-79
- tech/wink.md: Wink fa18-523-84
- tech/yarn.md: Yarn
- tech/zeromq.md: ZeroMQ fa18-523-79
- incomming/opennn.md: OpenNN
- incomming/orientdb.md: OrientDB hid-SP18-520
- incomming/pytorch.md: PyTorch
- incomming/rabbitmq.md: RabbitMQ
- incomming/redis.md: Redis
- incomming/tensorflow.md: TensorFlow

2 Preface



This collection of documents will help in creating cloud based clusters

2.1 Notation



If you click on the in a heading, you can go directly to the document in github that contains the next content. This is convenient to fix errors or make additions to the content.

\$

Content in bash is marked with verbatim text and a dollar sign

`$ This is a bash text`

[1]

References are indicated with a number and are included in the reference chapter [1]

Movies are indicated with a

Chapters marked with an are not yet complete or have some issue that we know about. These chapters need to be fixed.

Notes are indicated with a bulb and are written in italic and surrounded by bars

2.2 Format



The format of the entries are rather simple. Here is a sample entry:

```
## Technology Name
```

title	Technology Name
status	0
section	TBD
keywords	workflow, python

Here will be a text describing in meaningful but brief fashion what the technology is about and how it is used. References are included with brackets and the at sign as well as a label. The entry must be included in a bibliography file in the bib directory. you must make sure that the entry is not duplicated, either in the label or the content

[@hid-sp18-000-www-technology-name]. To avoid duplication in the label, we ask you to add your hid in case you take a class with us, or your github id in case you just like to contribute without taking a class.

```
> "This is a word by word quote from the  
> citation. Please note the > and the special  
> quotes. do not use a " to quote, as we use the begin  
> and end quote for formatting purposes"  
> [@hid-sp18-000-www-technology-name].
```

Please keep the following simple rules in mind. If we forgot a rule that could be helpful to you, let us know.

1. It is important that you do not just paste and copy, but that the entry you provide uses proper quotation rules and does not plagiarize. It is in your responsibility to assure this is the case. If you just paste and copy we rather like not to receive your contribution. Naturally if you are the author of the technology, you certainly can paste, but in case you do a class you will not be able to do so.
2. Any entry will be rejected if it uses the words below, above, left, and right to refer to descriptions in your text. In an Electronic publications terms usch as as below, and above make no sense.

3. Any figure that is copied from a source and does not have a citation and does not have a citation within the caption
4. Please create separate pull requests for each technology that you add or modify. We will reject all pull requests that try to fix multiple entries. THis is done in order to make reviews easier
5. The technology must be proceeded by ## and not just #
6. Do not forget that markdown as we use it requires to use proper quotes and not quotes that are used by editors such as Word. So use straight single and double quotes. To highlight a word use italics instead of quotes. Quotes are " and '. You can paste them form here and use.
7. When using references such as Web pages use a www- as prefix. Other prefixes are blog-, wiki-, youtube-.
8. An entry will be rejected if it does not have at least one citation.
9. An entry will be rejected if it uses urls in the text instead of citations.
10. A technology shoudl start with a sentence such as This TECHNOLOGYNAME is ... or similar. It must not start with a comment what bigdata is, or why bigdata is important. This can be looked up in our handbook instead.
11. use spell and grammar checker 12: You must eat least include the link to the main web page of the technology.
12. If a technology is retired please notify us and include the word retired in the keywords.

Once you have updated your technology, do not forget to put a :smiley: emoji in your title of the technology. This indicates to us that you like your technology to be reviewed. If it is still under construction use the emoji :construction:. This indicates to us that you have worked on it, but are not yet redy for review. In case you are seeing some issue with someones technology, you are allowed to mark it with the emoji :hand:. THis will indicate than that others may need to look at it. Also provide comments if you can.

The emojis are

ready to be reviewed and graded

ready to be reviewed and graded, same as

technology is picked by someone and is actively worked on

A reviewer suggests improvements

actively worked on by you

this entry needs improvements

do not remove an this indicates comments have been
made.

In case the entry needs to be rereviewed, please add a nother

No technologies will be reviewed if they have a or you need to
indicate this with a

2.3 Contributors



The document is located at

- <https://github.com/cloudmesh/technologies.git>

To add entries please use pull requests.

The following people have contributed to this document. If you find your name missing, please let us know or better create a pull request, and we add you. Please also notify us if you have not contributed so we can remove your name. We like to thank the following contributors in alphabetical order by lastname, firstname;

Abdul-Wahid, Badi; Abeykoon, Vibhatha; Agasti, Avadhoot; Agrawal, Saurabh; Ahmed, Tousif; Akurati, Niteesh Kumar; Anbazhagan, Karthik; Ardiansyah, Jimmy; Arnav, Arnav; Arora, Gagan; Athaley, Sushant; Ayna, Raveendran Gowtham; Balaga, Ajit; Balakrishnan, Abarnaa; Bays, Christopher; Bhatt, Himani; Carmickle, Ricky; Carmickle, Ricky; Castro, Andres; Chandwani, Nisha; Chandwani, Nisha; Chaturvedi, Dhawal; Chaudhary, Mrunal Lalitmohan; Chemburkar, Snehal Shrish; Chen, Huiyi; Cheruvu, Murali; Chheda, Chirag; Coulter, Cory; Deshmukh, Pravin; Devineni, Jyothi Pranavi; Dianprakasa, Arif; Dubey, Lokesh; Duffy, Kevin; Durbin, Matthew; Eliason, Neil; Fabianac, Tiffany; Fadnavis, Sarang; Gandavarapu, Harsh Reddy; Ganjoo, Ishita; Garner, Jeff; Gasiewicz, Robert; Guo, Wenting; Gupta, Abhishek; Han, Wenxuan; Heydarpour, Peyman; Hotz, Nicholas; Hotz, Nick; Huang, Yuan Ming; Irey, Ryan; Jain, Anurag Kumar; Jain, Pratik; Jones, Gabriel; Kagita, Mani Kumar; Kahn, Laura; Khamkar, Ajinkya; Khatiwada, Janaki; Kirzhner, Elena; Kodre, Vishwanath; Korrapati, Sahiti; Krishnakumar, Harshit; Kshirsagar, Hemant; Kugan, Uma; Kumar, Saurabh; Kumar,

Saurabh; Kuppuraj, Ashok; Langlois, J. Robert; Lawson, Matthew; Lingampalli, Anvesh Nayan; Lipe-Melton, Josh; Liu, Qiaoyi; Liu, Tony; Liu, Yuchen; Lu, Junjie; Mahendrakar, Mohan; Mallala, Bharat; Marks, Paul; Marni, Veera; Mathe, Dhanya; McClary, Scott; McCombe, Mark; Meier, Zachary; Merugureddy, Bhavesh Reddy; Methkupalli, Vasanth; Millard, Mathew; Miller, Ashley; Miller, Mark; Mishra, Govind; Mitchell, Jerome; Mwangi, Leonard; Naik, Abhishek; Neema, Ruchi Gupta; Ni, Juan; Nikolov, Dimitar; Niu, Geng; Parekh, Ronak; Patibandla, Brahmendra Sravan Kumar; Peterson, Tyler; Phillips, Judy; Raghatare, Rahul; Rai, Piyush; Ramachandran, Shahidhya; Ramanam, Srikanth; Ramaraju, Naveenkumar; Ravi, Anil; Ravi, Sowmya; Rawat, Neha; Roy, Budhaditya; Roy, Choudhury Sabyasachi; Rufael, Ribka; Russell, Peter; Sathe, Nandita; Satyam, Kumar; Schwartzer, Matthew; Shah, Vaibhav; Shane, Kevin; Sharma, Shridivya; Sharma, Yatin; Shen, Shiqi; Sheybani, Moghadam Saber; Shinde, Piyush; Shiverick, Sean; Simmons, Jordan; Singam, Ashok Reddy; Singh, Rahul; Sitharaman, Sriram; Sivaprasad, Sushmita; Smith, Michael; Sriramulu, Anand; Suri, Naren; Suryawanshi, Milind; Swargam, Prashanth; Sylla, Hady; Talebzadeh, Milani Mohammadreza; Thakkar, Bhavik Kamlesh; Thakre, Abhijit; Thompson, Tim; Tibenkana, Jacob; Townsley, Jeramy; Udojen, Nsikan; Unni, Sunanda; Usifo, Borga; Vegi, Karthik; Velayutham, Rahul; Venkatesan, Karthick; Vora, Sagar; Vujjini, Sri Megha; Vuppada, Ashok; Wang, Fugang; Wang, Jiaan; Wang, Weixuan; Wood, Ross; Wu, Yujie; Yadav, Diksha; Yang, Weipeng; Yezerets, Helen; Zhang, Miao; Zhu, Zhicheng; von Laszewski, Gregor;

Contributors are sorted by the first letter of their "combined Firstname and Lastname and if not available by their github ID. Please note that the authors are identified through git logs in addition to

some contributors added by hand. The git repository contains more than the documents included in this section. Thus not everyone in this list may have directly contributed to this document. However if you find someone missing that has contributed (they may not have used this particular git) please let us know. We will add you. The contributors that we are aware of include:

Abhishek Rapelli, Ankita Rajendra Alshi, Arnav, Averill Cate, Jr, Bertolt Sobolik, Bo Feng, Bo Li, Daniel hinders, Divya Rajendran, Fugang Wang, Geoffrey C. Fox, Gregor von Laszewski, Harika Putti, Hyungro Lee, IzoldalIU, Jatin Bhutka, Javier Diaz, Jay Stockwell, Juliette Zerick, Kelvin Liuwie, Miao Jiang, Min Chen, Nishad Tupe, Orly Esteban, Pavan Kumar Madineni, PavanMadineni, PrajaktaRPatil, Pramod Duvvuri, Pramod Kumar Duvvuri, Pulasthi Supun, Ravinder Lambadi, Ritesh Tandon, Ritu Sanjay, Saber Sheybani, Sahithya Sridhar, Sandeep Kumar Khandelwal, Selahattin AKKAS, Silvia Karim, Sohan Rai, Swarnima H. Sowani, Tim Whitson, Tyler Balson, Uma Kota, Vibhatha Abeykoon, Yuli Zhao, ahilgenkamp, aralshi, arijitsinha80, avheine, chmick, ckakara, ebbeall, fugangwang, harshadpitkar, janumudvari, jaystockwell, jeffliuhai, karankotz, manekbahl, mgm3IU, nhiyobie, nishadture, omkartamhankar, pulasthi, rdivyajd, sakkas20, saurabhIU, sushmitadash, umabkota, vbhoyar, wangton, yezi ma, yixihu, yulizhao, yushark91

2.4 Creating the Document

The documentation is very easy to create as it relies on pandoc. To install it you can do the following:

Mac OSX

Use homebrew

```
$ brew install pandoc  
$ brew install pandoc-citeproc
```

On Linux and Windows, please follow the directions for pandoc
Windows

Once you have installed pandoc you can create the book with our simple `Makefile` contained in the source directory. Simply clone the source and call make in the source dir

```
$ mkdir -p ~/github/cloudmesh  
$ cd ~/github/cloudmesh  
$ git clone https://github.com/cloudmesh/technologies.git  
$ cd technologies  
$ make todo  
$ make
```

To look at the book, open the text with your favorite e-book reader.
On OSX you can say

```
$ make view
```

...

3 Technologies

3.1 Accumulo



title	Accumul
status	90
section	NoSQL
keywords	NoSQL

Apache Accumulo, a highly scalable structured store based on Google's BigTable, is a sorted, distributed key/value store that provides robust, scalable data storage and retrieval. Accumulo is written in Java and operates over the Hadoop Distributed File System (HDFS), which is part of the popular Apache Hadoop project. Accumulo supports efficient storage and retrieval of structured data, including queries for ranges, and provides support for using Accumulo tables as input and output for MapReduce jobs. Accumulo features automatic load-balancing and partitioning, data compression and fine-grained security labels. Much of the work Accumulo does involves maintaining certain properties of the data, such as organization, availability, and integrity, across many commodity-class machines [2].

3.2 ActiveBPEL



title	ActiveBPEL
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Business Process Execution Language for Web Services (BPEL4WS or just BPEL) is an XML-based grammar for describing the logic to coordinate and control web services that seamlessly integrate people, processes and systems, increasing the efficiency and visibility of the business. ActiveBPEL is a robust Java/J2EE runtime environment that is capable of executing process definitions created to the Business Process Execution Language for Web Services. The ActiveBPEL also provides an administration interface that is accessible via web service invocations; and it can also be used to administer, to control and to integrate web services into a larger application [3].

3.3 ActiveMQ



title	ActiveMQ
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Apache ActiveMQ is a powerful open source messaging and Integration Patterns server [4]. It is a message oriented middleware (MOM) for the Apache Software Foundation that provides high availability, reliability, performance, scalability and security for enterprise messaging [5]. The goal of ActiveMQ is to provide standard-based, message-oriented application integration across as many languages and platforms as possible. ActiveMQ implements the JMS spec and offers dozens of additional features and value on top of this specifications. ActiveMQ is used in many scenarios such as heterogeneous application integration, as a replacement for RPC and to loosen the coupling between applications.

3.4 Aerobatic



title	Aerobatic
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Aerobatic is a platform that allows hosting static websites [6]. It used to be an ad-on for Bitbucket but now Aerobatic is transitioning to standalone CLI (Command Line Tool) and web dashboard. Aerobatic allows automatic builds to different branches. New changes to websites can be deployed using aero deploy command which can be executed from local desktop or any of CD tools and services like Jenkins, Codeship, Travis and so on. It also allows users to configure custom error pages and offers authentication which can also be customized. Aerobatic is backed by AWS cloud. Aerobatic has free plan and pro plan options for customers.

3.5 Agave



title	Agave
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Agave is an open source, application hosting framework and provides a platform-as-a-service solution for hybrid computing [7]. It provides everything ranging from authentication and authorization to computational, data and collaborative services. Agave manages end to end lifecycle of an application's execution. Agave provides an execution platform, data management platform, or an application platform through which users can execute applications, perform operations on their data or simple build their web and mobile applications [8].

Agave's API's provide a catalog with existing technologies and hence no additional appliances, servers or other software needs to be installed. To deploy an application from the catalog, the user needs to host it on a storage system registered with Agave, and submit to agave, a JSON file that shall contain the path to the executable file, the input parameters, and specify the desired output location. Agave shall read the JSON file, formalize the parameters, execute the user program and dump the output to the requested destination [7].

3.6 Airavata



title	Airavata
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Apache Airavata [9] is a software framework that enables you to compose, manage, execute, and monitor large scale applications and workflows on distributed computing resources such as local clusters, supercomputers, computational grids, and computing clouds. Scientific gateway developers use Airavata as their middleware layer between job submissions and grid systems. Airavata supports long running applications and workflows on distributed computational resources. Many scientific gateways are already using Airavata to perform computations (e.g. Ultrascan [10], SEAGrid [11] and GenApp [12]).

3.7 AllegroGraph



title	AllegroGraph
status	10
section	NoSQL
keywords	NoSQL

“AllegroGraph is a database technology that enables businesses to extract sophisticated decision insights and predictive analytics from their highly complex, distributed data that can't be answered with conventional databases, i.e., it turns complex data into actionable business insights” [13].

It can be viewed as a closed source database that is used for storage and retrieval of data in the form of triples (triple is a data entity composed of subject-predicate-object like Professor teaches students). Information in a triple store is retrieved using a query language. Query languages can be classified into database query languages or information retrieval query languages. The difference is that a database query language gives exact answers to exact questions, while an information retrieval query language finds documents containing requested information. Triple format represents information in a machine-readable format. Every part of the triple is individually addressable via unique URLs - for example, the statement Professor teaches students might be represented in RDF (Resource Description Framework). Using this representation, semantic data can be queried [14].

3.8 Amazon Dynamo



o: there should not be any
urls in the text

title	Amazon Dynam
status	10
section	NoSQL
keywords	NoSQL

Amazon explains DynamoDB as a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale [15]. It is a fully managed cloud database and supports both document and key-value store models. Its flexible data model and reliable performance make it a great fit for mobile, web, gaming, ad tech, IoT, and many other applications. DynamoDB can be easily integrated with big-data processing tools like Hadoop. It can also be integrated with AWS Lambda, an event driven platform, which enables creating applications that can automatically react to data changes. At present there are certain limits to DynamoDB. Amazon has listed all the limits in a web page titled Limits in DynamoDB [16].

3.9 Amazon Kinesis



title	Amazon Kinesis
status	10
section	Streams
keywords	Streams

Kinesis is Amazon's real time data processing engine. It is designed to provide scalable, durable and reliable data processing platform with low latency [17]. The data to Kinesis can be ingested from multiple sources in different format. This data is further made available by Kinesis to multiple applications or consumers interested in the data. Kinesis provides robust and fault tolerant system to handle this high volume of data. Data sharding mechanism in Kinesis makes it horizontally scalable. Each of these shards in Kinesis process a group of records which are partitioned by the shard key. Each record processed by Kinesis is identified by sequence number, partition key and data blob. Sequence number to records is assigned by the stream. Partition keys are used by partitioner (a hash function) to map the records to the shards i.e. which records should go to which shard. Producers like web servers, client applications, logs push the data to Kinesis whereas Kinesis applications act as consumers of the data from Kinesis engine. It also provides data retention for certain time for example 24 hours default. This data retention window is a sliding window. Kinesis collects lot of metrics which can be used to understand the amount of data being processed by Kinesis. User can use this metrics to do some analytics and visualize the metrics data. Kinesis is one of the tools part of AWS infrastructure and provides its users a complete software-as-a-service. Kinesis [18] in the area of real-time processing provides following key benefits: ease of use, parallel processing, scalable, cost effective, fault tolerant and highly available.



title	Amazon RDS
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Amazon Relational Database Service [19], also known as Amazon RDS, is a cloud service launched by Amazon to make users have a faster access to database, where users can save time of mandatory steps like setting up and operating. APIs of RDS and command-line tools are provided to users for

“access the full capabilities of a complete, self-contained MySQL 5.1 database instance in a matter of minutes” [20],

where original functionalities of MySQL are kept. For example, the processing power and storage space can be rescaled with an easy API call [20].

RDS also supports Multi-AZ deployments. In the deployment, RDS will keeps a hot-standby master, which would not generate replication lag so that data of the hot standby and the live master would always be the same. This kind of deployment helps keep data safe:

“In case of an instance failure, network outage, or even unavailability of the whole AZ of the master, the hot standby is automatically promoted to be the new master” [21].

Furthermore, there is a safety mechanism which prevents the case where hot standby fails: the point-in-time recovery. By using a consistent snapshot of data within data retention period, which can be extended to 8 days, the point-in-time recovery could help user

boot a database instance up to five minutes before the failure [21]. What else, besides those safety mechanisms, there is still another level of protection that users could use to give others permissions of access to their own RDS databases, which is called Identity and Access Management. Without permission from owner of the dataset, others would not be able to access the datasets [22].

3.11 Amazon Redshift fa18-423-06



title	Amazon Redshift
status	90
section	High level Programming
keywords	High level Programming

Amazon Redshift [fa18-423-06-www-Amazon-Redshift] is a data warehouse that is not only a columnar data store, but it uses massively parallel processing (MPP) [23]. Amazon boasts that the Redshift software is a fast, scalable, and cost-effective way to analyze and store data [24]. Most importantly, Amazon wants people to know that this software is fast and in the cloud. Unlike most data warehouses used in business, the Amazon Redshift software is constantly monitored by Amazon for a cheaper alternative to storing and accessing data.

In 2012, major data warehouse vendor ParAccel announced that Amazon had become a major investor in their company. Then, later in the year, Amazon came out with Amazon Redshift, which was very similar to the ParAccel technology [25]. The software is able to work without having to implement a new programming language.

"Just load up the cluster, connect your favorite query tool, and you're ready to go" [24].

The cluster that Amazon is referring to is the massively parallel processing that is the backbone of the data warehouse. MPP means using multiple processors and computers to perform computations in coordination ??? These clusters are the only way to move data through the data warehouse, a single node cannot be sent. Every cluster is made up of a leader, then nodes are in place behind that leader ??? These clusters are what gives Redshift the ability to query quickly in parallel. The structure of the data gives the Amazon

Redshift software the ability to be scaled up as the business grows, a major issue for businesses who grow quickly.

One of the major points that Amazon wants clients to know about Redshift is that the scalability, encryption, and maintenance is performed in the cloud by Amazon for a meager price of around \$1k per terabyte per year [24]. These factors are one of the main reasons that Redshift has been one of the fastest growing products in Amazon Web Service's portfolio ??? . Amazon's data warehouse service is sure to be one of the best and most convenient ways to store data, no matter the size of the business.

3.12 Amazon Route 53



title	Amazon Route 53
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Amazon Route 53 is a DNS (Domain Name System) service that gives developers and businesses a reliable way to route end users to Internet applications. The number 53 refers to TCP or UDP port 53, where DNS server requests are addressed [26].

When using Route 53 as your DNS provider, in case of a recursion, the query of fetching an IP address (of a website or application) always goes to the closest server location to reduce query latency. The Route 53 server returns the IP address enabling the browser to load the website or application. Route 53 can also be used for registering domain names and arranging DNS health checks to monitor the server [27].

3.13 Public Cloud: Amazon S3



title	Public Cloud: Amazon S3
status	10
section	File systems
keywords	File systems

Amazon Simple Storage Service (Amazon S3) is storage object which provides a simple web service interface to store and retrieve any amount of data from anywhere on the web [28]. With Amazon S3, users can store as much data as they want and can scale it up and down based on the requirements. For developers Amazon S3 provides full REST API's and SDK's which can be integrated with third-party technologies. Amazon S3 is also deeply integrated with other AWS services to make it easier to build solutions that use a range of AWS services which include Amazon CloudFront, Amazon CloudWatch, Amazon Kinesis, Amazon RDS, Amazon Glacier etc. Amazon S3 provides automatic encryption of data once the data is uploaded in the cloud. Amazon S3 uses the concept of Buckets and Objects for storing data wherein Buckets are used to store objects. Amazon S3 services can be used using the Amazon Console Management [29]. The steps for using the Amazon S3 are as follows: (1) Sign up for Amazon S3 (2) After sign up, create a Bucket in your account, (3) Create and object which might be an file or folder, and (4) Perform operations on the object which is stored in the cloud.

3.14 Public Cloud: Amazon SNS

52

fa18-523-



title	Public Cloud: Amazon SNS
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Amazon (Simple Notification Service) SNS is a secure messaging service that allows decoupling of distributed systems, microservices, and serverless applications. Amazons SNS allows for high-spanned messaging to a large number of endpoints that contributes to parallel processing such as Amazon SQS queues, HTTP/S webhooks, and more. SNS can also be used to notify users using SMS, email, or mobile push methods. Many users begin use of Amazon SNS by also using the AWS Management Console, AWS Command Line Interface (CLI), or AWS Software Development Kit (SDK).

All messages sent and received on Amazon SNS are stored on multiple servers and data centers. Amazon SNS is unique in that it is free and does not require the need for installation, upgrade, or other configuration. Amazon SNS also provides the ability to keep messages private and contained within an organization. This includes specific restrictions on who can and cannot publish or subscribe to certain topics. There are also configurations which allow filtering of messages of interest. For example, this allows only specific topic related messages to be included in the message thread instead of messages from every topic within the user's system. Low-level APIs are available for learners to use regarding the basics of building a topic thread.

Amazon SNS can be distributed in computer applications and data stores in business systems. A popular use for Amazon SNS is to distribute time-sensitive messages to users on mobile devices.

Amazon SNS has the advantage of relying on real-time applications to provide critical notifications. Specific Amazon SNS features include topic names that are limited to 256 characters, a unique Amazon Resource Name that can be referenced later, and an AWS ID [30].

3.15 Amazon



title	Amazon
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

ERROR: MANY BIBTEX ENTRIES MISSING

Amazon's AWS (Amazon Web Services) is a provider of Infrastructure as a Service (IaaS) on cloud. It provides a broad set of infrastructure services, such as computing, data storage, networking and databases. One can leverage AWS services by creating an account with AWS and then creating a virtual server, called as an instance, on the AWS cloud. In this instance you can select the hard disk volume, number of CPUs and other hardware configuration based on your application needs. You can also select operating system and other software required to run your application. AWS lets you select from the countless services. Some of them are mentioned next:

- Amazon Elastic Computer Cloud (EC2)
- Amazon Simple Storage Service (Amazon S3)
- Amazon CloudFront
- Amazon Relational Database Service (Amazon RDS)
- Amazon SimpleDB
- Amazon Simple Notification Service (Amazon SNS)
- Amazon Simple Queue Service (Amazon SQS)
- Amazon Virtual Private Cloud (Amazon VPC)

Amazon EC2 and Amazon S3 are the two core IaaS services, which are used by cloud application solution developers worldwide. [31]



title	Ambari
status	10
section	Monitoring
keywords	Monitoring

Ambari [32] is a console application based on web that is used to administrate Hadoop Clusters. It helps to manage, provide and monitor them. A Hadoop cluster is the place where one can store large amounts of unstructured data and use its computation ability to perform various analyzations on top of that data. Ambari helps in providing those clusters a help feature that automates and simplifies the process of preparing and deploying those clusters on different machines. It provides a business user a user-friendly dashboard to configure those services. It also provides a way to manage those clusters such as a centralized way to start those services and configure them again. It also helps in monitoring those clusters, where one can check the health condition of each cluster. If any node in a cluster goes down, or anything requires a special attention such as low disk space, or if anything requires an enhanced processing power, it sends an alert and notifies the administrator to take care of it [33].

It follows a client server architecture, where there exists an Ambari client and an Ambari Server. Ambari clients communicate with Ambari server using heartbeats through the interface of an Ambari Agent. Ambari server also exposes API to the users, so that one can customize the ambari application functionalities. Ambari server also maintains the database, where it stores the meta data [34]. Ambari Agent maintains an Agent Daemon which runs continuously and sends those heartbeats. In order to access the Ambari application, one uses security shell (ssh) to connect to the machine where it is installed and runs the commands such as Ambari-server start and

Ambari-client start. Once both are up and running, one can access the application on the web browser. On launching this application, one can manage different services in their Hadoop network [35]. One can also add other services as and when required. The dashboard provides information about live DataNodes, Memory Usage, NetworkUsage, CPU usage, HDFS Disk Usage, Cluster load and provides a summary report.



title	AMQP
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

AMQP stands for **Advanced Message Queueing Protocol**. It is an open standard source. It allows development of applications and work as middleware to broker messages between different processes, applications, or systems that need to talk to each other and pass on messages. It creates interoperability between clients and brokers. The protocol is binary, with features like negotiation, multichannel, portability, efficiency and asynchronous messaging [36].

It is commonly split into the below layers[37].

- Functional layer - Defines the commands for functioning on the part of the application
- Physical Layer - forms the physical (i.e. hardware) base for OSI to work.
- Data Link Layer - transfers the data between network nodes.
- Network Layer - directs the traffic (i.e. forwarding) between places.
- Transport Layer - Carry different techniques such as framing, channel multiplexing, data representation, etc., between the server and the application.
- Session Layer - responsible of managing the session between applications.
- Presentation (Syntax) Layer - working to shape and present the data to be processed.

- Application Layer - setting and ensuring common grounds - reaching the applications - for communication. (This is where AMQP lives!)

Application layer is the one with which user interacts with.

3.17.1 Advantages

- Rapid and guaranteed message deliveries [37].
- Reliability and message acknowledgments
- Globally share and monitor updates and also to enable communication between different systems that are connected
- Full synchronous functionality for systems as well as improved reliability

3.17.2 Components:

The major components of AMQP are: Exchange, message queue and the bindings. The **exchange** is the part of the broker that receives and routes the messages to the **queue**. Bindings are the rules for distributing messages from exchange to the queues.

In AMQP, message brokers translate to applications which receive the actual messages and route (i.e. transfer) them to relevant parties. After receiving the message from the client, the exchange process them and route them to the queue.

Following are types of exchanges [37]:

1. Direct Exchange

Direct exchange type involves the delivery of messages to queues based on routing keys. Routing keys can be considered as additional data defined to set where a message will go. Typical use case for direct exchange is load balancing tasks in a round-robin way between workers.

2. Fanout Exchange

Fanout exchange completely ignores the routing key and sends any message to all the queues bound to it.

Use cases for fanout exchanges usually involve distribution of a message to multiple clients for purposes similar to notifications:

Sharing of messages (e.g. chat servers) and updates (e.g. news)
Application states (e.g. configurations)

3. Topic Exchange

Topic exchange is mainly used for pub/sub (publish-subscribe) patterns. Using this type of transferring, a routing key alongside binding of queues to exchanges are used to match and send messages.

Whenever a specialized involvement of a consumer is necessary (such as a single working set to perform a certain type of actions), topic exchange comes in handy to distribute messages accordingly based on keys and patterns.

4. Headers Exchange

Headers exchange constitutes of using additional headers (i.e. message attributes) coupled with messages instead of depending on routing keys for routing to queues.

Being able to use types of data other than strings (which are what routing keys are), headers exchange allow differing routing mechanism with more possibilities but similar to direct exchange through keys.

3.18 Ansible



title	Ansible
status	10
section	DevOps
keywords	DevOps

Ansible is an IT automation tool that automates cloud provisioning, configuration management, and application deployment [38]. Once Ansible gets installed on a control node, which is an agentless architecture, it connects to a managed node through the default OpenSSH connection type [39].

As with most configuration management softwares, Ansible distinguishes two types of servers: controlling machines and nodes. First, there is a single controlling machine which is where orchestration begins. Nodes are managed by a controlling machine over SSH. The controlling machine describes the location of nodes through its inventory.

Ansible manages machines in an agent-less manner. Ansible is decentralized, if needed, Ansible can easily connect with Kerberos, LDAP, and other centralized authentication management systems.

3.19 Any2Api



title	Any2Api
status	10
section	DevOps
keywords	DevOps

This framework [40] allows user to wrap an executable program or scripts, for example scripts, chef cookbooks, ansible playbooks, juju charms, other compiled programs etc. to generate APIs from your existing code. These APIs are also containerized so that they can be hosted on a docker container, vagrant box etc Any2Api helps to deal with problems like scale of application, technical expertise, large codebase and different API formats. The generated API hide the tool specific details simplifying the integration and orchestration different kinds of artifacts. The APIfication framework contains various modules:

1. Invokers, which are capable of running a given type of executable for example cookbook invoker can be used to run Chef cookbooks
2. Scanners, which are capable of scanning modules of certain type for example cookbook scanner scans Chef cookbooks.
3. API impl generators, which are doingthe actual work to generate the API implementation.

The final API implementation [41] is is packages with executable in container. The module is packaged as npm module. Currently any2api-cli provides a command line interface and web based interface is planned for future development. Any2Api is very useful for by devops to orchestrate open source ecosystem without dealing with low level details of chef cookbook or ansible playbook or puppet.

It can also be very useful in writing microservices where services talk to each other using well defined APIs.

3.20 Apache Ant



title	Apache Ant
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Ant is a Java library and command-line tool whose mission is to drive processes described in build files as targets and extension points dependent upon each other. The main known usage of Ant is the build of Java applications. Ant supplies a number of built-in tasks allowing to compile, assemble, test and run Java applications. Ant can also be used effectively to build non Java applications, for instance C or C++ applications. More generally, Ant can be used to pilot any type of process which can be described in terms of targets and tasks. Ant is written in Java. Users of Ant can develop their own

"antlibs containing Ant tasks and types, and are offered a large number of ready-made commercial or open-source antlibs. Ant is extremely flexible and does not impose coding conventions or directory layouts to the Java projects which adopt it as a build tool. Software development projects looking for a solution combining build tool and dependency management can use Ant in combination with Apache Ivy. The Apache Ant project is part of the Apache Software Foundation [42].



title	Apache Apex
status	10
section	Hadoop, NoSQL
keywords	Hadoop, NoSQL

Apache Apex [43] is a YARN (Hadoop 2.0) native platform that unifies cloud and batch processing . This project was developed under Apache License 2.0. It can be used for processing both streams of data and static files making it more relevant in the context of present day internet and social media. It is can be used to leverage the present Hadoop platform and make it easy for developers to learn and build a streaming application intuitively. This allows the users to focus on writing their application logic without mixing operational and functional concerns. The platform handles application execution, dynamic scaling, state checkpointing and recovery, etc. It enables the code reusage by not having to make drastic changes to existing application code and also provides interoperability using its technology stack.

An application may consist of one or more operators each of which define some logical operation to be done on the tuples arriving at the operator. These operators [44] are connected together to form a network of streams. A streaming application is represented by a DAG that consists of operators and streams. The Apex platform comes with support for web services and metrics. This enables ease of use and easy integration with current data pipeline components. DevOps teams can monitor data in action using existing systems and dashboards with minimal changes, thereby easily integrating with the current setup. With different connectors and the ease of adding more connectors, Apache Apex easily integrates with an existing dataflow [45].

Another component of this technology stack is Apex Malhar [46] which provides a library of connectors and logic functions. It provides connectors to existing file systems, message systems and relational, NoSQL and Hadoop databases, social media. It also provides a library of compute operators like Machine Learning, Stats and Math, Pattern Matching, Query and Scripting, Stream manipulators, Parsers and UI and Charting operators [47].



title	Apache Arrow
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Arrow [48] is an open-source standardized memory format first designed to solve the complexity of data exchange between two systems started in early 2016 according to one of the project initiator Wes Mckinney who is also the project initiator of Pandas [49]. One of the problems Arrow solves is the cost of serialization when data transfers between systems. The most common way to exchange data between systems is to export data into a CSV file or a JSON file first then import into another system, which is also an inefficient way that has high serialization cost. The solution Arrow offers is a standardized data specification for different systems, a columnar memory format for dataframes. The columnar memory-layout specifies the format that how in-memory data is stored.

“The layout is highly cache-efficient in analytics workloads and permits SIMD optimizations with modern processors” [50].

This standard memory format is supported by many systems including: Calcite, Cassandra, Drill, Hadoop, HBase, Ibis, Impala, Kudu, Pandas, Parquet, Phoenix, Spark, and Storm [48]. It is cost efficient exchanging data between these systems because they all support Arrow’s memory format, and resource will not be wasted on serialization and deserialization. The Apache Arrow Homepage emphasize the flexibility of Apache Arrow because it supports “a wide variety of industry-standard programming languages” [48]. This means programming languages like Java, Python, C are all supported by Apache Arrow. Besides the columnar memory-layout, one other

feature was adopted by Apache Arrow to support efficient data sharing. According to Mckinney, Plasma object was donated to Apache Arrow in 2017 from UC Berkeley RISELab.

"It keeps track of how many processes have a reference to a particular dataset" [49].

So when all processes are done with the dataset, Plasma will empty the memory of the dataset to for others uses.



title	Apache Beam
status	10
section	MapReduce, Batch and streaming
keywords	MapReduce, Batch and streaming

One of the biggest obstacles in working with big data is integrating the various frameworks, APIs and SDKs. To tackle this problem, Google along multiple others came up with an integrative model that coalesces multiple data workflows like batch, interactive and streaming and also acts as a solitary platform for cloud as well as local development. This level of integrability allows users to switch between technologies seamlessly [51].

A unified model is offered by Apache Beam for both outlining and executing huge information-oriented work processes inside an information preparing, information mix, and information ingestion according to the Apache Beam venture page [52]. The language-specific SDKs are supported by Apache Beam till date are Java, Python, Go, Scala (as Scio). Apache Apex, Apache Flink, Apache Gearpump, Apache Samza , Apache Spark, Google Cloud Dataflow are some of the distributed processing backends which can be used with Runners supported by Beam. Assume you have a MapReduce and now you have to join these occupations with Spark which needs heaps of effort and cost. After this, the exertion and cost you have to change to another stage need to refactor your employments once more. A layer is offered by data stream between the code & job runtime. A unified model is allowed by the SDK for executing your information handling logic with the assistance of Dataflow SDK that keeps running on distinctive backends. There is no compelling reason to refactor or change the code any longer.

In the Apache Beam SDK, there are four builds [53]

- Pipelines: There are couple of calculations like information, output, and handling.
- PCollections: Pipeline I/o data. Represent logical sets of data but don't contain data
- PTransforms: (data processing step) I/O are more than one PCollections. Take collections as input and produce collections as outputs.
- I/O Sources and Sinks APIs: To read and write data to pipeline.

Advantages of using Beam is it has an open ecosystem, it's community driven, vendor independent. One can program pipelines in any language that has beam SDK. Multiple Beam runners are available in the market. There is no vendor lock-in i.e., one can run any language on any runner in beam, there is no language lock-in either i.e., there's no need to stick to a particular language beforehand. One can use transforms to convert from any language.



title	Apache Derby
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Apache Derby is an Apache database sub-project designed based on java, SQL and JDBC standards [54]. > “Apache Derby (previously distributed as IBM Cloudscape) is a > relational database management system (RDBMS) developed by the > Apache Software Foundation that can be embedded in Java programs and > used for online transaction processing” [55].

Relational database systems have long been popular among developers to store data. Its popularity stems from the fact that information between various tables can be linked together using keys that uniquely identify any atom in a table. Not to mention that RDBMS provide easy to manage data. Examples of RDBMS include MS SQL [56], Oracle [57], Derby [54] etc.

Derby is implemented completely in Java. Devices that make use of the Java Micro edition can take full advantage of Derby, given the fact that it only leaves about 2MB as footprint for both the embedded JDBC driver and the base engine. It requires no maintenance (until the application changes), and hence can be embedded in applications written in Java, where details are hidden from the user. Derby runs on most OSs including windows, AIX, solaris, UNIX and Mac OS.

In the famous book Apache Derby - Off to the Races by Zikopoulos, Baklarz, and Scott (2005), the authors are of the view that

“not all client/server or Web applications require the muscle of an enterprise-class infrastructure database”

[58].

Only about 20-30% of applications actually require RDBMS capabilities. Furthermore, their hosting environments usually do not have the system requirements to run full-fledged data engines. Note that these systems still need robustness and scalability to ensure data integrity and this is where Apache derby comes in. Another, benefit of using Derby is that it eliminates the use of a database administrator. The DB can be managed programmatically from the application itself [58].

The following scenarios describe instances where Derby could be the suitable choice:

1. Small business client database applications: Eliminating the need for a DBA can significantly reduce costs
2. Local registries and repositories: Since the DB is fully transactional, developers need not worry about crashes that can destroy configuration files.
3. Small business client/server and Web-based applications: To build websites low on maintenance but high on reliability. This ensures that businesses have plenty of head room for seasonal spikes.

3.25 Apache Flex



title	Apache Flex
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Flex [59] is an open source application framework for building and maintaining mobile and web applications that deploy consistently on multiple browsers, desktops and mobile devices. It was initially developed by Macromedia and then acquired by Adobe Systems. It was later donated to the Apache Software Foundation in 2011 [60]. It can pull data from multiple back-end sources such as Java, Spring, PHP, Ruby, .NET, Adobe ColdFusion, and SAP and display it visually allowing users to drill down into the data for deeper insight and even change the data and have it automatically updated on the back end [61].



title	Apache - HAWK
status	90
section	High level Programming
keywords	High level Programming

Apache HAWQ is a Hadoop SQL database built for massively parallel processing. The database offers advanced parallel processing combined with machine learning tools and the scalability of Hadoop [62]. The architecture of Apache HAWQ aims to provide low latency query responses that can scale to datasets as large as a petabyte [62]. This results a reduction of time needed to explore large datasets and implement complex machine learning models. The improved query speed is also beneficial for consumer applications that rely on learning models and big data. Another benefit of Apache HAWQ is its compatibility with SQL based applications as the database is ANSI SQL compliant. Business Intelligence and data visualization tools are also compatible allowing easy exploration of large datasets with increased speed. The compatibility and similarity allows a user to leverage familiar skills to work with the database and achieve the desired results [62].

Apache HAWQ is native to Hadoop which allows for easy scaling of nodes to handle capacity or performance requirements [62]. The tool also provides easy integration of Apache MADlib machine learning libraries that can be used to work with data in Apache HAWQ. As of Aug 15, 2018 Apache HAWQ was approved as a top level project for the Apache Software Foundation [62]. As momentum increases new features and bug fixes are planned to be implemented in order to improve the tool.

The project is continuing to move forward but in February of 2018 the organization Pivotal decided to end the availability of HAWQ (Pivotal

HDB) [63]. In the early design of HAWQ Pivotal was a key player in the development through the creation of the Pivotal Greenplum enterprise database [62]. Now the project will continue as an open source project.

3.27 Apache Knox



title	Apache Knox
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

According to [64],

“the Apache Knox Gateway is a REST API Gateway for interacting with Apache Hadoop clusters.”

REST stands for Representational State Transfer and is web architectural style designed for distributed hypermedia systems and defines a set of constraints. [65] API Gateways manage concerns related to

“Authentication, Transport Security, Load-balancing, Request Dispatching (including fault tolerance and service discovery), Dependency Resolution, Transport Transformations.” [66]

Although every Apache Hadoop cluster has its own set of REST APIs, Knox will represent all of them as

“a single cluster specific application context path.”

[64]

Knox protects Apache Hadoop clusters, by way of its gateway function, by aiding

“the control, integration, monitoring and automation of critical administrative and analytical needs.” [64]

Some Apache Hadoop Services that integrate with Knox are,

“Ambari, WebHDFS (HDFS), Templeton (Hcatalog), Stargate (Hbase), Oozie, Hive/JDBC, Yarn RM, [and] Storm.” [64]

Apache Knox has a configuration driven method to aid in the addition of new routing services. [64] This allows support for new and custom Apache Hadoop REST APIs to be added to the Knox gateway quickly and easily. [64] This technology would be best placed under the interoperability category.



title	Apache OODT
status	10
section	Data System Framework
keywords	Data System Framework

OODT [67] stands for Object Oriented Data Technology which is funded initially by NASA's Office of Space Science. OODT is an open source architecture and provided by the Apache Software Foundation. Apache OODT is a data system framework that was built with a vision to share data across heterogeneous and distributed data repositories. It focusses primarily on big data processing and information integration.

"It is both an architecture as well as a framework that enables data production, data discovery, data distribution, data analysis and data access on a large scale" [68].

OODT is very user-friendly. It is very easy to manage the workflow and execute tasks even for a non-programmer. Its processing unit consists of shell scripts that are customizable depending on the workflow. The main components include a file manager, a workflow manager and a resource manager which each manage different modules of the data workflow [67].

OODT additionally takes into consideration remote execution of employments on versatile computational foundations so that computational and data-intensive concentrated handling can be coordinated into OODT's data processing pipelines using cloud computing and high-performance computing environments which are intrinsic to OODTs framework [69].

Conventional data processing pipelines comprise custom UNIX shell scripts and delicately composed code while Apache OODT employs organized XML-based capturing of data processing pipeline that can be comprehended easily by non-developers to create, alter and smooth workflow and task execution. OODT spans across different disciplines and enables interoperability among information skeptic frameworks in any field. Apache OODT makes use of Lattice Computing and Apache Mesos to perform various computing operations across distributed systems in addition to utilizing cloud computing services. This enables developers to build highly dispersed, versatile and scalable data platforms that can process even enormous volumes of data effortlessly [70].

3.29 Apache Ranger



title	Apache Ranger
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Ranger [71] is open source software project designed to provide centralized security services to various components of Apache Hadoop. Apache Hadoop provides various mechanism to store, process and access the data. Each Apache tool has its own security mechanism. This increases administrative overhead and is also error prone. Apache Ranger fills this gap to provide a central security and auditing mechanism for various Hadoop components [72]. Using Ranger, Hadoop administrators can perform security administration tasks using a central UI or Restful web services. He can define policies which enable users/user-groups to perform specific action using Hadoop components and tools. Ranger provides role based access control for datasets on Hadoop at column and row level. The blog article [73] explains that the row level filtering and dynamic data masking are most important features of Apache Ranger. Ranger also provides centralized auditing of user acces and security related administrative actions.

3.30 Apache Tomcat



title	Apache Tomcat
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache tomcat is an open source java servlet container. [74] It is used in IT industry as a HTTP web server which listens to the requests made by web client and send responses. The main components of tomcat are cataline, coyote and jasper. The most stable version of Apache Tomcat server is version 8.5.11. Apache tomcat is released under Apache License version 2. [75] As it is cross platform, it can run in any platform or OS like Windows, UNIX, AIX or SOLARIS etc. It is basically an integral part of many java based web application.

3.31 Apatar



title	Apatar
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration



title	appfog
status	100
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

AppFog can be described as

“a platform as a service (PaaS) provider” [76].

Platform as a service (PaaS) is a cloud computing model in which a third-party provider delivers hardware and software tools – usually those needed for application development – to users over the internet. A PaaS provider hosts the hardware and software on its own infrastructure. As a result, PaaS frees users from having to install in-house hardware and software to develop or run a new application [77].

PaaS does not typically replace a business's entire IT infrastructure. Instead, a business relies on PaaS providers for key services, such as application hosting or Java development. A PaaS provider builds and supplies a resilient and optimized environment on which users can install applications and data sets. Users can focus on creating and running applications rather than constructing and maintaining the underlying infrastructure and services [78].

The Centurylink AppFog is a public multi-tenant Platform-as-a-Service (PaaS) based on Cloud Foundry, that makes deploying scalable, robust, high performing cloud-native apps fast and easy for developers. AppFog supports common developer application runtime that include Java, Node.js, PHP, Python, Go, Ruby, Static Websites and ASP.NET. Integrating 3rdParty Add-ons such as database, messaging, load balancing, monitoring and more are easily provisioned through

the AppFog Marketplace[79].

AppFog is a Platform as a Service that can be integrated on-premise into a company's data center. It is also available as a public service. The company was originally founded as PHPFog before changing its name early in 2011 after receiving \$8 million in funding. The company began focusing more on a private PaaS strategy. AppFog competes in a crowded market that includes Pivotal's Cloud Foundry, Red Hat's OpenShift, ActiveState's Stackato and Apprenda.

Heroku and Engine Yard are two of the leaders in the public PaaS market [80].

Cloudfoundry.org pushed the boundaries of platform-as-a-service by providing an open source tool set that developer could use. From developer point of view, AppFog has many infrastructures, letting you choose them and give you portability of those infrastructure, as well as code [81].

CenturyLink has acquired AppFog to strengthen its subsidiary Savvis. The acquisition will enhance the Savvis Platform-as-a-Service and the data hosting solutions [82].

AppFog is simply setup Cloud Foundry CLI (Command Line Interface) to target an AppFog region, then push your applications up to AppFog within seconds. Accomplished without any operational overhead, so developers can focus on applications without operational or infrastructure distractions [79]. Incoming HTTP/HTTPS request route traffic to an application via platform managed Load Balancing clusters, with no configuration required by system user. Cloud Foundry application instance provisioning is automatically balanced across AppFog regional data centers. All application data and configuration parameters are automatically synced between the provisioned application instances across data centers in a region. Cross data center balancing increases the application's availability [79].

3.33 AppScale



title	AppScale
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

AppScale is an application hosting platform. This platform helps to deploy and scale the unmodified Google App Engine application, which run the application on any cloud infrastructure in public, private and on premise cluster [83]. AppScale provide rapid, API development platform that can run on any cloud infrastructure. The platform separates the app logic and its service part to have control over application deployment, data storage, resource use, backup and migration. AppScale is based on Google's App Engine APIs and has support for Python, Go, PHP and Java applications. It supports single and multimode deployment, which will help with large, dataset or CPU. AppScale allows to deploy app in thee main mode i.e. dev/test, production and customize deployment [84].

3.34 Askalon



title	Askalon
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Askalon was developed at the University of Innsbruck [85]. It is application development as well as a runtime environment. It allows easy execution of distributed work flow applications in service oriented grids. It uses a Service Oriented Architecture. Also, for its Grid middleware it uses the Globus Toolkit. The work flow applications are developed using Abstract Grid Work flow Language (AGWL). The architecture has various components like the resource broker responsible for brokerage functions like management and reservation, information service for the discovery and organization of resources and data, metascheduler for mapping in the Grid, performance analysis for unification of performance monitoring and integration of the results and the Askalon scheduler.

The Metascheduler is of special significance since it consists of two major components - the workflow converter and the scheduling engine. The former is responsible for conversion of traditional workflows into directed acyclic graphs (DAGs) while the later one is responsible for the scheduling of workflows for various specific tasks. It has a conventional pluggable architecture which allows easy integration of various services. By default, the Heterogeneous Earliest Finish Time (HEFT) is used as the primary scheduling algorithm.



title	Atmosphere
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Atmosphere [86] or as it was previously known as iPlant Collaborative at the time when it was launched is a Cloud Computing platform with a web interface developed by CyVerse with the aim of providing high power computational resources to everyone in the research community to analyze and perform various resource heavy tasks on possible real-world complex datasets and enable cutting edge research and data-driven discovery. The platform also educates people new to Cloud Computing with various tutorials available free. It is being managed by the funds from the National Science Foundation (NSF). It gives the user to create a Virtual Machine in the Cloud. The concept of Virtual Machines (VMs) [87] has become very popular among developers since it gave them the opportunity to work with multiple Operating Systems (OS) on a single machine which was usually their local machine or their personal computer. With the power of Cloud Computing, CyVerse aimed to create such virtual machines (VMs) pre-configured for various domain specific tasks and capable of running data and computationally intensive algorithms in the cloud and make it accessible for the developers or people in the field of research who wanted machines with such capability. To use Atmosphere, the user can use his local machine, since everything is done in the Cloud there are no minimum local machine requirements to use Atmosphere. This service can be accessed from any device and is entirely platform independent. But prior knowledge of a Linux environment or command-line is useful and necessary to a certain extent [88]. The user first needs to register with CyVerse and create an account. The user then needs to request access to Atmosphere to utilize the computer power in the cloud. Atmosphere is built in a way

that makes it easy for multiple users or researchers working on a single project easy. This is done by enabling screen sharing between multiple users, so people can work collaboratively on data analysis. The team responsible has also built a web portal for easy access to the resources in the cloud and continue to keep improving the platform to accommodate the needs of the research community in a timely manner.

3.36 Avro



title	Avr
status	90
section	Message and Data Protocols
keywords	Message and Data Protocols

Apache Avro is a data serialization system, which provides rich data structures, remote procedure call (RPC), a container file to store persistent data and simple integration with dynamic languages [89]. Avro depends on schemas, which are defined with JSON. This facilitates implementation in other languages that have the JSON libraries. The key advantages of Avro are schema evolution - Avro will handle the missing/extra/modified fields, dynamic typing - serialization and deserialization without code generation, untagged data - data encoding and faster data processing by allowing data to be written without overhead.

AWS Elastic Beanstalk fa18-423-03



title	AWS Elastic Beanstalk
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Elastic Beanstalk was developed by Amazon Web Services (AWS) ??. This platform was developed to allow developers to create and test their applications inside a single program, rather than leveraging multiple different platforms. Overall, Elastic Beanstalk provides a scalable and so-called elastic route to developing an application. However, because all branches of development are held within Elastic Beanstalk, it can be difficult to navigate for the first time [90]. The

basic architecture is relatively straightforward, however,

" It uses Amazon Elastic Compute Cloud (EC2) instances, Amazon Simple Storage Service (S3) buckets, and load balancers to manage your application architecture for you" [91].

AWS offers 750 hours of t2.micro EC2 time per month for free, but adding servers is not free. A helpful feature to developers is that Elastic Beanstalk features rolling updates. This allows developers to make updates to the application while the application is still online. However, this is not the case for changes to the environment. These changes must be committed to Git first and then pushed. Elastic Beanstalk offers a few functions that automatically work with applications such as provisioning, load balancing, autoscaling, and application health monitoring [92]. Elastic Beanstalk has an open architecture, thus the developer can deploy applications not written in Web language, however for web developers, Java, Node.js, PHP, Python, Ruby, and .Net are supported. Continuing on with the scalability of the service, although the toolkit is completely available for developers to use and AWS Elastic Beanstalk is completely free to use (within the specified hours per month), the

"resources used to store and run their applications"

are what need to be paid for [92].

3.37 AWS OpsWorks



title	AWS OpsWorks
status	90
section	DevOps
keywords	DevOps

AWS Opsworks is a configuration service provided by Amazon Web Services that uses Chef, a Ruby and Erlang based configuration management tool [93], to automate the configuration, deployment, and management of servers and applications. There are two versions of AWS Opsworks. The first, a fee based offering called AWS OpsWorks for Chef Automate, provides a Chef Server and suite of tools to enable full stack automation. The second, AWS OpsWorks Stacks, is a free offering in which applications are modeled as stacks containing various layers. Amazon Elastic Cloud Compute (EC2) instances or other resources can be deployed and configured in each layer of AWS OpsWorks Stacks [94].

3.38 Azure Blob



title	Azure Blob
status	10
section	File systems
keywords	File systems

Azure Blob storage is a service that stores unstructured data in the cloud as objects/blobs. Blob storage can store any type of text or binary data, such as a document, media file, or application installer. Blob storage is also referred to as object storage. The word 'Blob' expands to Binary Large OBject [95]. There are three types of blobs in the service offered by Windows Azure namely block, append and page blobs [96]. 1. Block blobs are collection of individual blocks with unique block ID. The block blobs allow the users to upload large amount of data. 2. Append blobs are optimized blocks that helps in making the operations efficient. 3. Page blobs are compilation of pages. They allow random read and write operations. While creating a blob, if the type is not specified they are set to block type by default. All the blobs must be inside a container in your storage. Azure Blob storage is a service for storing large amounts of unstructured object data, such as text or binary data, that can be accessed from anywhere in the world via HTTP or HTTPS. You can use Blob storage to expose data publicly to the world, or to store application data privately. Common uses of Blob storage include serving images or documents directly to a browser, storing files for distributed access, streaming video and audio, storing data for backup and restore, disaster recovery, and archiving and storing data for analysis by an on-premises or Azure-hosted service. Azure Storage is massively scalable and elastic with an auto-partitioning system that automatically load-balances your data. Blob storage is a specialized storage account for storing your unstructured data as blobs (objects) in Azure Storage. Blob storage is similar to existing general-purpose storage accounts and shares all the great durability, availability, scalability, and

performance features. Blob storage has two types of access tiers that can be specified, hot access tier, which will be accessed more frequently, and a cool access tier, which will be less frequently accessed. There are many reasons why you should consider using BLOB storage. Perhaps you want to share files with clients, or off-load some of the static content from your web servers to reduce the load on them [95].



title	Azure Data Factory
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Azure Data Factory is a tool that can be used to process and manage large data flows generated by applications, sensors and other data sources [97]. The process by which Azure Data Factory does this is by collecting data across many sources, organizing the data, publishing data and monitoring the data flow [97]. This is achieved through four major components. The first component of Azure Data Factory is to connect data sources. Second, once the data is collected to a central data store in the cloud it can be transformed for use in business analytics or other applications. This transformation is done through the use of tools such as HDInsight Hadoop, Spark, Data Lake Analytics, and Machine Learning [97]. The third component of Azure Data Factory is the ability to publish your transformed data so that business intelligence tools and other applications can utilize the data. Finally, Azure Data Factory provides a portal to monitor the data pipeline that has been set up.

The advantage of Azure Data Factory described on the Microsoft website is the 70 data connectors that are available and can be implemented with clicks, not code, in the user interface [98]. There are also many other ETL tools on the market that have similar functionality. Choosing an ETL tool is an important decision when creating a pipeline to manage your data. Some important factors to consider are the ability to connect to the sources that you are using to collect data, the speed of development, stability and cost [99]. Azure Data Factory has a robust list of data connectors. However, there are other ETL tools such as MuleSoft that offer an ecosystem of data connectors which extends the list of data sources that can be

connected out of the box [100]. The development speed can be a challenge for many of the ETL tools available. This depends on your data sources as well as the complexity of the transform process. Azure Data Factory has an extensive list of functionality but it may be difficult to implement depending on what is needed. Azure is one of the leaders in data storage and processing. As a more established player in the space, the reliability of the platform may be greater than some of Microsoft's smaller competitors. As for pricing the cost is comparable to the cost of Amazon Web Services Glue which is a similar ETL tool for AWS services [101] [102].

In summary Azure Data Factory is an enterprise level ETL tool that supports the collection of data, transformation process, publication of data and monitoring of data flows. There are many ETL tools available and choosing the right one depends on many factors. Azure Data Factory is a good tool for larger businesses that may already have a presence on Azure but are also generating data from a variety of other sources.

3.40 Azure Machine Learning



title	Azure Machine Learning
status	90
section	Application and Analytics
keywords	Application and Analytics

Azure Machine Learning is a cloud based service that can be used to do predictive analytics, machine learning or data mining. It has features like in-built algorithm library, machine learning studio and a web service [103]. In built algorithm library has implementation of various popular machine learning algorithms like decision tree, SVM, linear regression, neural networks etc. Machine learning studio facilitates creation of predictive models using graphical user interface by dragging, dropping and connecting of different modules that can be used by people with minimal knowledge in the machine learning field. Machine learning studio is a free service for basic version and comes with a monthly charge for advanced versions. Apart from building models, studio also has options to do preprocessing like clean, transform and normalize the data. Webservice provides option to deploy the machine learning algorithm as ready to consume APIs that can be reused in future with minimal effort and can also be published.

3.41 Azure Queues



title	Azure Queues
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Azure Queues storage is a Microsoft Azure service, providing inter-process communication by message passing [104]. A sender sends the message and a client receives and processes them. The messages are stored in a queue which can contain millions of messages, up to the total capacity limit of a storage account [105]. Each message can be up to 64 KB in size. These messages can then be accessed from anywhere in the world via authenticated calls using HTTP or HTTPS. Similar to the other message queue services, Azure Queues enables decoupling of the components [106]. It runs in an asynchronous environment where messages can be sent among the different components of an application. Thus, it provides an efficient solution for managing workflows and tasks. The messages can remain in the queue up to 7 days, and afterwards, they will be deleted automatically.



title	Azure SQL
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Azure SQL is Microsoft's integrated cloud base relational database service [107]. It provides Platform as a service (PaaS) and Infrastructure as a service (IaaS). It is simply SQL server on Azure but in a more intelligent way.

The SQL Server based cloud service enables IT professionals and developers to store structured data on the cloud that can be highly scalable as per the business needs without any need of infrastructures. This enables to build data driven websites and applications on top of it. It is an intelligent relational cloud database service providing broad SQL Server engine compatibility. It manages Database with minimum administration cost and in a highly secure way. Also, it provides SQL server that lowers the cost of buying SQL server license [108].

Azure SQL offers some interesting features to set up, scale and operate relational database on the cloud. Some of them are Geo-replication that allows to replicate whole database in a completely different region, Dynamic data masking that enables to make sensitive data safe from a restricted user, Automatic tuning which increases performance of executed queries automatically through indexing and history logs, automatic back-ups, supports recovery up to any point, data encryption at rest also secures data from unexpected threats [109]. Azure SQL has reporting portal on cloud which supports various methods to create, view and share business reports [110].

Azure SQL is capable of handling almost anything related to databases but there are a few limitations. For instance, we cannot check If SQL agent is running or not and T- SQL support is also limited [107]. Azure SQL pricing model is based on elastic pool and managed instances. Deploying Microsoft Azure SQL database allows a secure user level access and has also reduced IT infrastructure and storage costs remarkably, simplifying workload management, thereby delivering improved analytics performance.

3.43 Azure Stream Analytics



title	Azure Stream Analytics
status	10
section	Streams
keywords	Streams

Azure Stream Analytics is a platform that manages data streaming from devices, web sites, infrastructure systems, social media, internet of things analytics, and other sources using real-time event processing engine [111]. Jobs are authored by

“specifying the input source of the streaming data, the output sink for the results of your job, and a data transformation expressed in a SQL-like language.”

Some key capabilities and benefits include ease of use, scalability, reliability, repeatability, quick recovery, low cost, reference data use, user defined functions capability, and connectivity. Available documentation to get started with Azure Stream Analytics [112]. Azure Stream Analytics has a development project available on github [113].



title	Public Cloud: Azure Table
status	10
section	NoSQL
keywords	NoSQL

Azure Table [114] storage facility developed by Microsoft is one of most advanced cloud storages which can cater to all the requirements of big data [115]. Azure storage technology has certain salient features particularly conducive for big data computation. Azure tables are massively scalable, which is essential to meet the requirements of today's application processes. These cloud storage capabilities are equipped to endure transient hardware failures, which makes them highly reliable and secure. It can be used to replicate redundant data across geographies in order to safeguard against any natural catastrophe. The maintenance of these cloud servers is also managed by the Microsoft authorities, making it easy for the users to protect themselves from any critical problems. The data sets kept in these cloud storage facilities can be accessed from anywhere over http or https. Such an alternative assures that the users can access the data stored in Azure Table storage facility when they need it. Azure table can be used by clients in any field to protect and back up their data at any scale which is required. Moreover, the cloud is highly reliable and cost efficient being easy to use and maintain. All these features of Azure Table storage make it one of the most popular cloud storage facilities.

3.45 Azure



title	Azure
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Microsoft Corporation (MSFT) markets its cloud products under the Azure brand name. At its most basic, Azure acts as an infrastructure-as-a-service (IaaS) provider. IaaS virtualizes hardware components, a key differentiation from other -as-a-service products. IaaS

“abstracts the user from the details of infrastructure like physical computing resources, location, data partitioning, scaling, security, backup, etc” [116].

However, Azure offers a host of closely-related tool and products to enhance and improve the core product, such as raw block storage, load balancers, and IP addresses [114]. For instance, Azure users can access predictive analytics, Bots and Blockchain-as-a-Service as well as more-basic computing, networking, storage, database and management components [117] [114]. The Azure website shows twelve major categories under Products and twenty Solution categories, e.g., e-commerce or Business SaaS apps.

Azure competes against Amazon’s Amazon Web Service, even though IBM (SoftLayer and Bluemix) and Google (Google Cloud Platform) offer IaaS to the market [118]. As of January 2017, Azure’s datacenters span 32 Microsoft-defined regions, or 38 declared regions, throughout the world [114].



title	Berkeley DB
status	10
section	NoSQL
keywords	NoSQL

Berkeley DB [119] is an open source data management library. Because it is open source, anyone can use it for free. It has a few features that make it highly versatile for developers. First, the database library is scalable. Berkeley DB's initial library is

“under 300 kilobytes of text space on common architecture, but it can manage databases up to 256 terabytes in size” [120].

Second, it exhibits an application programming interface that allows the user to enter a simple call function to operate the database libraries and management services. Berkeley DB supports C, C++, Java, Tcl, PHP, Python, and Perl. Furthermore, developers can embed Berkeley DB directly into their program and with the API, data store, concurrent, transactional, and replication configuration options all run on the back end.

Berkeley was originally released in 1996. It was developed by Margo Seltzer and Keith Bostin of Sleepycat Software and sold to Oracle Corporation in 2006. Since the database management library software's inception it has gone through many advances and iterations. When first released, Berkeley DB was designed only to handle large data, greater than the size of the hash buckets being used, and to provide

“constant time mapping between values and page addresses” [121].

One of the most complex developments was introducing a recovery manager in Berkeley DB 2.0.6. Berkeley DB also offers Btree, Queue, Recno, and Hash access methods. As mentioned previously, there is a library interface because the developers realized that the service required front end functionality for both the applications and the internal code. Furthermore, the developers introduced managers for each of the configuration options that required all their own architectures.

3.47 Bioconductor fa18-xxx-xx



title	Bioconductor
status	10
section	Application and Analytics
keywords	Application and Analytics

Bioconductor is an open source and open development platform used for analysis and understanding of high throughput genomic data. Bioconductor is used to analyze DNA microarray, flow, sequencing, SNP, and other biological data. All contributions to Bioconductor are under an open source license. The goals of Bioconductor

“include fostering collaborative development and widespread use of innovative software, reducing barriers to entry into interdisciplinary scientific research, and promoting the achievement of remote reproducibility of research results” [122].

The Bioconductor is primarily based on R, as most components of Bioconductor are released in R packages [123]. Extensive documentation is provided for each Bioconductor package as vignettes, which include task-oriented descriptions for the functionalities of each package. Bioconductor has annotation functionality to associate

“genomic data in real time with biological metadata from web databases such as GenBank, Entrez genes and PubMed.”

Bioconductor also has tools to process genomic annotation data.

3.48 BioKepler



title	BioKepler
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

BioKepler is a Kepler module of scientific workflow components to execute a set of bioinformatics tools using distributed execution patterns [124]. It contains a specialized set of actors called

"bioActors" for running bioinformatic tools, directors providing distributed data-parallel (DPP) execution on Big Data platforms such as Hadoop and Spark they are also configurable and reusable [125]. BioKepler contains over 40 example workflows that demonstrate the actors and directors [126].



title	BitTorrent
status	10
section	Data Transport
keywords	Data Transport

BitTorrent [127] refers to that method of data transfer which allows large data files to be broken down into smaller ones for efficient data sharing. The smaller files with the data can be downloaded from several sources simultaneously. It is indeed the most common protocol that enables peer to peer file sharing with ease and most efficiently.

This technology works by allowing users to create a small file, referred to as torrent, which contains information about the files and the computers that manage the distribution of those files. To download a file, one needs to find and open the torrent created and start downloading files piece by piece. The peers can also share the pieces with any other interested party trying to download the same files. This sharing enhances easy file downloads as more pieces of data can be found in multiple sources. BitTorrent provides faster download speeds at a significantly low bandwidth as a result of breaking down files into smaller pieces.

BitTorrent does not only provides faster download speeds but also solves the problem of high bandwidth needed. In the traditional download mode, the file is generally transferred from the server to the client. Which means more people download, more bandwidth needed. But BitTorrent breaks the file to the small piece, it uses a pyramid scheme to achieve sharing. Which means it's not only one server, every client how has to download the same thing will be the server, no super high bandwidth needed.



o: try to start with a sentence that you
wrote and not a quote

title	Blaze
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

grammar

3.50.1 Old text

Blaze library translates NumPy/Pandas-like syntax to data computing systems (e.g. database, in-memory, distributed-computing). This provides Python users with a familiar interface to query data in a variety of other data storage systems. One Blaze query can work across data ranging from a CSV file to a distributed database.

Blaze presents a pleasant and familiar interface regardless of what computational solution or database we use (e.g. Spark, Impala, SQL databases, No-SQL data-stores, raw-files). It mediates the users interaction with files, data structures, and databases, optimizing and translating the query as appropriate to provide a smooth and interactive session. It allows the data scientists and analyst to write their queries in a unified way that does not have to change because the data is stored in another format or a different data-store. [128]

3.50.2 New text

Blaze is a series of tools which are designed to make the operation of data more convenient for users.

“The Blaze ecosystem is a set of libraries that help

users to store, describe, query and process data" [129].

It is designed for Python users, with an ecosystem to allow Python users process to large-scale computation, as we know, computation of huge amount data is not easy to finish efficient, but Blaze can use its library to help Python user to do that, the ecosystem could be very useful when doing the computation.

To help Python users speed up the computation of large-scale data, the ecosystem could make the computation simpler than before. Since Blaze provides a unique language, data shape, that could describe data, and it is out of the data processing. And there is also a common interface to request data, which is also out of the data processing. And a uniform utility library to transfer data, which is called odo. The desk is a parallel computational engine.

But Blaze has nothing to do with the computation, the number operation is also conducted in other systems. Traditional SQL, or new technology Spark, or the Pandas package in Python. But for Blaze, it does not use the calculate packages in Python, nor interacting with libraries.

Blaze is good at operating a little part of data which belongs to an larger dataset, although the volume maybe be small, but due to its small-scale, the process speed could be very quick so Blaze could be powerful when dealing with the data.

All in all, Blaze is a high-level user interface for all Python users who are interesting in operating dataset and conduct the computation. There is a symbolic expression system for the requests of data and make it clear to the database, there is also a translator for the requests for many different databases. It could be very useful when conducting a multi-platform project.

Due to the special structure of Blaze, it could allow a piece of code to run well in many different backends, which could be very convenient for developers. And if Blaze could be applied widely, Python users with different computing tools could work together with less

pressure.

3.51 Blazegraph



title	Blazegraph
status	10
section	NoSQL
keywords	NoSQL

Blazegraph is a graph database also supporting property graph, capable of clustered deployment. A graph database is a NoSQL database. It is based on a graph theory of nodes and edges where each node represents an element such as user or business and each edge represents relationship between two nodes. It is mainly used for storing and analyzing data where maintaining interconnections is essential. Data pertaining to social media is best example where graph database can be used.

Blazegraph's main focus is large scale complex graph analytics and query. The Blazegraph database runs on graphics processing units (GPU) to speed graph traversals. [130]

Lets now see how Blazegraph handles data. [131] Blazegraph data can be accessed using REST APIs.

Blazegraph supports Apache TinkerPop, which is a graph computing framework.

For graph data mining, Blazegraph implements GAS (Gather, Apply, Scatter) model as a service.



title	BlinkDB
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

BlinkDB [132] is a large-scale data storage system based on platforms spark and shark that is designed to be compatible with Apache Hive. It is a huge system built for executing smart and intelligent SQL queries on extensive volumes of information. The Blink Data Base enables its users to compromise slightly on query accuracy for faster response times there by enabling interactive users to execute their queries on data samples. The Blink Data Base provides results to these queries run on data sample with meaningful errors rather than running tedious queries on massive data which is a tedious process. The BlinkDB achieves this using two techniques namely -

- Employing an adaptive optimization framework that randomly samples and creates multidimensional data from original data which can be used to run sample queries in shorter time durations.
- Using a dynamic selection methodology that estimates and chooses an appropriate test dataset based on query's precision requirement as well as reaction time necessities [133].

The BlinkDB acts as an extension to the Apache Hive System by including two noteworthy parts to it. They are -

- An online inspecting module that creates and maintains samples over time.
- A run-time test sample selection module that creates an Error-Latency-Profile (ELP) for queries.

To make decisions on what samples to create, QCSs (A tool for Querying, Clustering and Summarizing) that appear in queries are used. Once the samples are created, disseminated repository sampling or binomial inspecting systems are used to make a scope of uniform and stratified tests over various measurements. The EPL or Error-Latency-Profile created during the run-time describes the rate which mistake reaction time declines and is employed to choose the random sample that best stratifies the client's requirements. BlinkDB is basically a distributed sampling-based estimated querying framework that endeavors to achieve a superior balance between efficiency and other desired properties of querying. As a result of such flexibilities, it allows users to pose aggregation queries over stored data in addition to ensuring shorter response times as well as error bound constraints.

3.53 Blueprints



title	Blueprints
status	10
section	DevOps
keywords	DevOps

In [134], it is explained that

“IBM Blueprint has been replaced by IBM Blueworks Live.”

In [135], IBM Blueworks Live is described

“as a cloud-based business process modeller, belonging under the set of IBM SmartCloud applications”

that as [136] states

“drives out inefficiencies and improves business operations.”

Similarly to Google Docs, IBM Blueworks Live is

“designed to help organizations discover and document their business processes, business decisions and policies in a collaborative manner.”

While Google Docs and IBM Blueworks Live are both simple to use in a collaborative manner, [135] explains that IBM Blueworks Live has the

“capabilities to implement more complex models.”

3.54 Boto



title	Bot
status	90
section	DevOps
keywords	DevOps

The latest version of Boto is Boto3 [137]. Boto3 is the Amazon Web Services (AWS) Development Kit (SDK) for Python [138]. It enables the Python developers to make use of services like Amazon S3 and Amazon EC2 [139]. It provides object oriented APIs along with low-level direct service [140]. It provides simple in-built functions and interfaces to work with Amazon S3 and EC2.

Boto3 has two distinct levels of APIs - client and resource [139]. One-to-one mappings to underlying HTTP API is provided by the client APIs. Resource APIs provide resource objects and collections to perform various actions by accessing the attributes. Boto3 also comes with 'waiters'. Waiters are used for polling status changes in AWS, automatically. Boto3 has these waiters for both the APIs - client as well as resource.

3.55 Buildstep



title	Buildstep
status	90
section	DevOps
keywords	DevOps

Buildsteps is an open software developed under MIT license. It is a base for Dockerfile and it activates Heroku-style application. Heroku is a platform-as-service (PaaS) that automates deployment of applications on the cloud. The program is pushed to the PaaS using git push, and then PaaS detects the programming language, builds, and runs application on a cloud platform [141]. Buildstep takes two parameters: a tar file that contains the application and a new application container name to create a new container for this application. Build script is dependent on buildpacks that are prerequisites for buildstep to run. The builder script runs inside the new container. The resulting build app can be run with Docker using docker build -t your_app_name command. [142].

3.56 Caffe



title	Caffe
status	90
section	Application and Analytics
keywords	Application and Analytics

Caffe is a deep learning framework made with three terms namely expression, speed and modularity [143]. Using Expressive architecture, switching between CPU and GPU by setting a single flag to train on a GPU machine then deploy to commodity cluster or mobile devices. Here the concept of configuration file will comes without hard coding the values. Switching between CPU and GPU can be done by setting a flag to train on a GPU machine then deploy to commodity clusters or mobile devices.

It can process over 60 million images per day with a single NVIIA k40 GPU. It is being used bu academic research projects, startup prototypes, and even large-scale industrial applications in vision, speech, and multimedia.



o: quotation preferred

title	Cascading
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Amount of data required to be processed for big data applications that help organizations is huge and is increasing day by day. Hadoop is used to help process such large amount of data. Hadoop does this by dividing data among multiple machines in a cluster and then allowing applications to perform needed task using MapReduce strategy. Cascading comes into picture by allowing users to create these applications that perform operations to process raw data, create workflows that streamline functions working on processed data and finally export results that are generated by performing MapReduce per functions created in earlier stages [144].

While working with big data, we can start by working on small sub-set of data on local machine and test if the application is working as intended. If satisfied, the application can then be scaled to work on big data. Other way Cascading is useful is that programmers can use other languages besides Java to work with big data by developing their API in language they are good at using. Programmers are typically good at thinking about how to perform filters, joins on big data but not so in terms of MapReduce. Cascading allows to create applications where program is first thought of in these terms that come naturally. Cascading allows to visualize whole process and hence we can understand where system is running into issues and fix those. This also helps figure out cost of performing certain tasks and optimize cost/time.

Other big advantage of cascading is scalability. Applications

developed using Cascading can be scaled quickly to more and different data for it to work and perform desired data processing. Cascading works with other platforms such as Tez and Flink. It can also be used locally. Cascading is used by many companies that need big data handling such as Twitter, LinkedIn, eBay, Nokia to build their large-scale deployments using Cascading as basic building block [144].

3.58 Cassandra



title	Cassandra
status	10
section	NoSQL
keywords	NoSQL

Apache Cassandra is an open-source distributed database management for handling large volume of data across commodity servers [145]. It works on asynchronous masterless replication technique leading to low latency and high availability. It is a hybrid between a key-value and column oriented database. A table in cassandra can be viewed as a multi dimensional map indexed by a key. It has its own

"Cassandra Query language (CQL) query language for data extraction and mining. One of the demerits of such structure is it does not support joins or subqueries. It is a java based system which can be administered by any JMX compliant tools."

3.59 CDAP



title	CDAP
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

CDAP [146] stands for Cask Data Application Platform. CDAP is an application development platform using which developers can build, deploy and monitor applications on Apache Hadoop. In a typical CDAP application, a developer can ingest data, store and manage datasets on Hadoop, perform batch mode data analysis, and develop web services to expose the data. They can also schedule and monitor the execution of the application. This way, CDAP enables the developers to use single platform to develop the end to end application on Apache Hadoop.

CDAP documentation [147] explains the important CDAP concepts of CDAP Dataset, CDAP Application and CDAP Services. CDAP Datasets provide logical abstraction over the data stored in Hadoop. CDAP Applications provide containers to implement application business logic in open source processing frameworks like map reduce, Spark and real time flow. CDAP applications also provide standardize way to deploy and manage the apps. CDAP Services provide services for application management, metadata management, and streams management. CDAP can be deployed on various Hadoop Platforms such as Apache Hadoop, Cloudera Hadoop, Hortonworks Hadoop and Amazon EMR. CDAP sample apps [148] provide explain how to implement apps on CDAP platform.

3.60 CDF



title	CDF
status	10
section	File management
keywords	File management

Common Data Format is a conceptual data abstraction for storing, manipulating, and accessing multidimensional data sets [149]. CDF differs from traditional physical file formats by defining form and function as opposed to a specification of the bits and bytes in an actual physical format.

CDF's integrated dataset is composed by following two categories: (a) Data Objects - scalars, vectors, and n-dimensional arrays. (b) Metadata - set of attributes describing the CDF in global terms or specifically for a single variable [150].

The self-describing property (metadata) allows CDF to be a generic, data-independent format that can store data from a wide variety of disciplines. Hence, the application developer remains insulated from the actual physical file format for reasons of conceptual simplicity, device independence, and future expandability. CDF data sets are portable on any of the CDF-supported platforms and accessible with CDF applications or layered tools. To ensure the data integrity in a CDF file, checksum method using MD5 algorithm is employed [151].

Compared to HDF format, CDF permitted cross-linking data from different instruments and spacecraft in ISTP with one development effort [152]. CDF is widely supported by commercial and open source data analysis/visualization software such as IDL, MATLAB, and IBM's Data Explorer (XP).



title	CDMI
status	10
section	Interoperability
keywords	Interoperability

CDMI or Cloud Data Management Interface is an industry standard proposed by the Storage Networking Industry Association (SNIA)[153], which defines the protocol for applications to access, create, retrieve, update, delete and administer data on cloud for application developers[154]. The Interface mostly utilizes the RESTful principles in its design. It helps client realize the features cloud storage offers and further aids in managing the containers and data present in them. Additionally, metadata of the data system can also be accessed and linked with containers, so that it can be used to in differentiating the data treatments. Types of metadata used by CDMI include HTTP, storage system, user metadata, etc[155].

CDMI offers an interface to manage data and as well as one to store and access it. Data path is the interface by which storage and retrieval of data can be done by CDMI while control path is means by which data is managed by the CDMI. CDMI can also manage cloud storage properties when other data path interfaces are used too [156].

Capabilities are the configuration parameters that are provided by compliant implementations which are mostly Boolean values or numerical values providing data about the system. The files in file system with their user-defined metadata together are called objects which are uniquely identified by the object id or names, the sizes of these objects vary from system to system based on its limit. These objects are then placed in container hierarchy [155].

"Users and groups created in a domain share a

common administrative database and are known to each other on a "first name" basis, i.e. without reference to any other domain or system" [154].

Currently, CDMI can be implemented by most of the cloud storage applications with help of an adapter if they have a proprietary interface or directly with the current interfaces. The administrative applications also use the interface to manage containers, domains, security access, and monitoring/billing information, even for storage that is functionally accessible by legacy or proprietary protocols [156].

3.62 Celery



title	Celery
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

"Celery is an asynchronous task queue/job queue based on distributed message passing. The focus of celery is mostly on real-time operation, but it equally scheduling. In celery there are execution units, called tasks, are executed concurrently on a single or more worker servers using multiprocessing, Eventlet, or gevent. Tasks can execute asynchronously (in the background) or synchronously (wait until ready). Celery is easy to integrate with web framework. Celery is written in python whereas the protocol can be implemented in any language" [157].

"Celery is a simple, flexible, and reliable distributed system to process vast amounts of messages, while providing operations with the tools required to maintain such a system" [158].



o: several paragraphs
have no backing ref

title	Ceph
status	10
section	File systems
keywords	File systems

Ceph [ref missing] is an open source storage platform which supports object, block and file system storage in one unified system. The system is highly scalable and thousands of clients can access a large amount of data easily. It uses commodity hardware to run. A Ceph Storage cluster can consist of a very large number of nodes. Nodes communicate each other to redistribute the data to increase throughput.

Ceph is based on RADOS which is a short version of Reliable Autonomic Distributed Object Store [ref missing]. LIBRADOS [ref missing] which is a library gives direct access to applications. LIBRADOS [ref missing] supports C, C++, Java, Python, Ruby, and PHP. Also using RADOS Gateway (RADOSGW) [ref missing], data in the storage can be accessed from Amazon S3 and OpenStack Swift.

The storage cluster receives data from clients and stores data as objects. Each object stored on an Object Storage Device (OSD) [ref missing]. These devices store data as objects in a flat namespace. Directory hierarchies are not used in these devices. Each object in an OSD has an ID, data in binary format and metadata [159].

Ceph Block Storage provides users to mount Ceph as a provisioned block device. Ceph block devices stripe the data across the cluster. It also integrates with KVM. Thus, KVM's have virtually unlimited storage [160].

Ceph also provides a traditional file system storage. Files are mapped to the objects in the cluster. Clients can mount the filesystem as a kernel and use it.

Ceph uses CRUSH algorithm (Controlled Replication Under Scalable Hashing) that decides how data will be stored and retrieved by the storage locations [159].

Ceph has four main parts:

Monitors(ceph-mon): a Ceph monitor keeps track of the cluster. They maintain the maps. These maps are required for coordination between daemons and clients.

Managers(ceph-mgr): a Ceph manager keeps track of cluster's current state and many metrics like storing utilization.

Ceph Object Storage Daemon (OSDs, ceph-osd): OSDs are responsible for storing data. OSDs make replications and rebalancing. Also, they regularly send heartbeats to other daemons.

Metadata Server(MDS - ceph-mds): It stores the metadata information [161].



title	Chef
status	10
section	DevOps
keywords	DevOps

Chef [ref missing] is an open source configuration management tool and automation platform for Devops. This tool is written in Ruby [ref missing] and Erlang [ref missing] and utilizes Ruby DSL [ref missing] for writing system configurations. The goal of the tool is to reduce the amount of manual, repetitive tasks that are required when performing infrastructure management or deploying new machines. Chef also has the capability to integrate with any cloud technology. Chef is both able to provide assistance in deployment and management of servers in the cloud and for in-house applications [162].

Chef is made up of several basic building blocks. Some examples of these building blocks are Recipes, Cookbooks, Resources, Attributes, Files, Templates, and Metadata.rb [162]. While these are some of the basic building blocks, the main three components are: The Recipe, The Cookbook, and the Resource. Recipes are utilized to set up an infrastructure node, they determine what should be installed, files that should be written, etc. Recipes are the workhorse of the Chef tool. By combining several recipes together, a cookbook is created. This provides oversight and easy deployment of several recipes at the same time for many nodes on the system. Finally, a resource is the basic component of a Recipe. Resources are tweaked to what the user needs set up on each of the nodes. Several resource selections come together to make up a Recipe.

Chef's architecture is divided between three main components. These are the Chef Workstation, Chef Server, and Chef Nodes. The Chef

Workstation is where the configurations are developed prior to deployment. The Chef Server is where the configurations are deployed once they have been finalized. The Chef Server organizes and plans how Nodes will be set up and organized. Finally, the Chef Nodes are the machines that are managed by the Chef Server through Recipes, Cookbooks, etc. Combining all these components together makes Chef an extremely powerful tool for managing and automating Devops tasks.

3.65 Cinder



title	Cinder
status	90
section	File systems
keywords	File systems

“Cinder is a block storage service for Openstack” [163].

Openstack Compute uses ephemeral disks meaning that they exist only for the life of the Openstack instance i.e. when the instance is terminated the disks disappear. Block storage system is a type of persistent storage that can be used to persist data beyond the life of the instance. Cinder provides users with access to persistent block-level storage devices. It is designed such that users can create block storage devices on demand and attach them to any running instances of OpenStack Compute [164]. This is achieved through the use of either a reference implementation (LVM) or plugin drivers for other storage. Cinder virtualizes the management of block storage devices and provides end users with a self-service API to request and consume those resources without requiring any knowledge of where their storage is actually deployed or on what type of device [163].

3.66 CiNET



title	CINET
status	10
section	Application and Analytics
keywords	Application and Analytics

A representation of connected entities such as

“physical, biological and social phenomena”

within a predictive model [165]. Network science has grown its importance understanding these phenomena Cyberinfrastructure is middleware tool helps study Network science,

“by providing unparalleled computational and analytic environment for researcher” [166].

Network science involves study of graph a large volume which requires high power computing which usually cant be achieve by desktop. Cyberinfrastructure provides cloud based infrastructure (e.g. FutureGrid) as well as use of HPC (e.g. Shadowfax, Pecos). With use of advance intelligent Job mangers, it select the infrastructure smartly suitable for submitted job.

It provides structural and dynamic network analysis, has number of algorithms for

“network analysis such as shortest path, sub path, motif counting, centrality and graph traversal”.

CiNet has number of range of network visualization modules. CiNet is actively being used by several universities, researchers and analysist.

3.67 Cisco Intelligent Automation for Cloud

title	Cisco Intelligent Automation for Cloud
status	90
section	DevOps
keywords	DevOps

Cisco Intelligent automation for cloud desires to help different service providers and software professionals in delivering highly secure infrastructure as a service on demand. It provides a foundation for organizational transformation by expanding the uses of cloud technology beyond its infrastructure [167]. From a single self-service portal, it automates standard business processes and sophisticated data center which is beyond the provision of virtual machines. Cisco Intelligent automation for cloud is a unified cloud platform that can deliver any type of service across mixed environments [168]. This leads to an increase in cloud penetration across different business and IT holdings. Its services range from underlying infrastructure to anything-as-a-service by allowing its users to evaluate, transform and deploy the IT and business services in a way they desire.



title	Cloud Foundry
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Cloud Foundry is an open source platform as service (PaaS) cloud application available through public and private cloud distributions developed initially in-house at VMware and mainly written in languages such as Java, Ruby, and Go [169].

In today's era, many organizations are facing the challenge of migrating their applications to the cloud and decommissioning the old infrastructure. It is a challenging task to make applications cloud native, means unaware of the underlying infrastructure that cloud services are providing. Cloud Foundry shines by providing the platform that makes application agnostic of infrastructure and runs smoothly leveraging cloud resources thereby reducing development cycle runtime [170].

Cloud Foundry's dedicated subsystems such as BOSH, CF cloud controller, and Router help to serve and scale the applications online flexibly [171].

From the computing quality perspectives, Cloud Foundry is optimized to deliver performance, scalability, availability, resilience; and support multi-tenant compute efficiencies for fast application development and deployment. Cloud Foundry Supports many languages such as Ruby, Java, Scala, Node.js, Python, and provides flexibility to deploy them on multi-cloud IaaS environments that includes OpenStack, AWS, and VSphere. It also allows developers to run their code on multiple database services such as MySQL, Postgres, MongoDB, Redis, RabbitMQ. The open source PaaS is highly customizable which

makes deploying and scaling applications fast and secure [172].

The Cloud Foundry architecture includes seven core groups of components which provide a platform for application deployment and lifecycle management, integration to other services such as databases or third-party SaaS providers ,and application execution [171].

Since the Cloud Foundry is an open source project; it is available from the Cloud Foundry Foundation as well as from a variety of software providers as a product or a service. The Cloud Foundry foundation built BOSH deployment system to allow Cloud Foundry to interact with the underlying infrastructure. The BOSH system allows deploying software over multiple VM's and also performs monitoring, failure recovery, and patching with zero to minimal downtime [169].

Security is a critical aspect, especially within a shared environment. The designated cloud resources may have application accessed by multiple consumers or companies. Cloud Foundry mitigates security threats by using measures such as isolation to customer data and containerized application, encryption techniques, role-based access control, and monitoring of resource starvation to prevent the possible security attack [171].

To summarize, Cloud Foundry provides a cloud computing platform to run the applications in the secure environment with self-healing capacity, centralized management, and ease of integration and maintenance [172].



title	Cloudability
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Cloudability[fa18-523-86-www-cloudability]. is the first financial management tool for monitoring and analyzing every cloud expense across any organization. It brings transparency to how and where organizations spend money on cloud resources, giving them the power to reap the most value from cloud usage possible [fa18-523-86-www-cloudability]. Cloudability has some competitors for the cloud cost management business such as Cloud Cruiser, Rightscale and CloudCheckr. The tool can send budget alerts and suggestions via SMS and email, as well as an API to connect cloud billing and usage data to other business or financial systems. The user of tool is primarily Cloudability cloud administrators, finance teams, and other corporate or IT professionals. Cloudability can track multiple cloud providers and help companies analyses their spending flows and make their cloud computing more flexible, avoiding unnecessary expenses and hidden costs.

It supports vendors such as Amazon Web Services (AWS). Cloudability's AWS analytics tools make it easier than ever to track, manage and communicate your AWS spending and usage. Create any detailed AWS spending and usage report you need with a simple point-and-click analytics interface.[fa18-523-86-www-cloudability-AWS]. Cloudability also has an AWS Reserved Instance Planning tool that guides users through the Reserved Instance purchase process, Users can view cost reports for all vendors, and will receive a budget alert when the cost is close to a predefined spending limit.

Recently Microsoft Azure was added into the support list, Cloudability

can assign multiple users based on projects, departments, and more. However, account group functionality is only available in enterprise services. Cloudability provided two pricing plan packages: Pro and Enterprise. Both plans are based on customer request demand model, the monthly fee is based on the amount of cloud cost monitored by the user. And more, for Enterprise pricing package, it comes with additional features and higher levels of support from Cloudability[fa18-523-86-www-cloudability].

3.70 CloudBees



title	CloudBees
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Cloudbees provides Platform as a Service (PaaS) solution, which is a cloud service for Java applications [173]. It is used to build, run and manage the web applications. It was created in 2010 by Jenkins. It has a continuous delivery platform for DevOps, and adds a enterprise-grade functionality with an expert level support. Cloudbees is better than the traditional Java platform as it requires no provision of the nodes, clusters, load balancers and databases. In cloudbees the environment is constantly managed and monitored where a metering and scale updating is done on a real time basis. The platform ships with verified security and enhancements assuring less risk for sharing sensitive information. It implies the task of getting the platform accessed by every user using the feature Jenkins Sprawl [174].



title	Cloudmesh
status	10
section	Interoperability
keywords	Interoperability

##Old

Cloudmesh client allows to easily manage virtual machines, containers, HPC tasks, through a convenient client and API. Hence cloudmesh is not only a multi-cloud, but a multi-hpc environment that allows also to use container technologies. Cloudmesh is currently developed as part of classes taught at Indiana University.

##New

Cloudmesh encompasses a method which uses a client to manage virtual clusters pertaining to big data sets. Cloudmesh combines various resources together to deliver a holistic application that allows for users to experiment and leverage big data applications. FutureSystems, Amazon Web Services, Azure, and HP Cloud are just a few examples of cloudmesh systems. Cloudmesh can function as an HPC (High Performance Computing) environment by providing a simple hosted or standalone interface as well as Python and iPython APIs [175]. The goal here is to effectively work with batch queues. The Cloudmesh interface includes a new concept called Rain. This new concept is fully integrated within Cloudmesh and provides a mechanism by which specific users can access projects depending on their roles [176]. This is a pertinent feature that allow for users to restrict access to specific projects. Rain includes the ability to complete any provisioning directly on the server, including any type of OS provisioning. The concept is entitled bare metal provisioning [176].

There are three layers that comprise the cloudmesh architecture are described below:

"The three layers of the Cloudmesh architecture include a Cloudmesh Management Framework for monitoring and operations, user and project management, experiment planning and deployment of services needed by an experiment, provisioning and execution environments to be deployed on resources to (or interfaced with) enable experiment management, and resources" [177].

Depending on your application needs, Cloudmesh can be set up to run either as a single host on one machine, or as a multi-tenancy solution. There are three primary services that start up allowing for the application to run effectively; Cloudmesh Database, Cloudmesh Web Service, and Cloudmesh Task Service.

"Cloudmesh Database: A NOSQL database in which we record which virtual machines run on which IaaS. This allows us to have a federated view of the heterogeneous clouds. Cloudmesh Web Service: Provides a Graphical user interface to manage virtual machines and HPC tasks Cloudmesh Task Service: As cloudmesh is a multi user systems many tasks need to be handles in parallel. To achieve this we are using an AMPQ queue and coordinate the execution of managing multiple virtual machines for multiple users" [177].

Cloudmesh has proven to work very effectively in academic settings by allowing instructors and student to collaborate together on assignments and projects. Cloudmesh allows users to leverage all of their cloud based computing accounts in one location. The FutureSystems application, leveraged by students from Indiana University, is an example of a Big Data cloudmesh environment that allows students to post projects, share project ideas, and interact with other students and instructors. It has proven to be and effective

educational tool in regards to STEM classes as it allows for immediate feature and transparency of everyone's work and code in order to provide a true learning environment.

3.72 CloudML



title	CloudML
status	90
section	DevOps
keywords	DevOps

CloudML a research project initiated by SINTEF in 2011 [178]. Cloud computing facilitates to shared and virtualized computer capabilities like storage, memory, CPU, GPU and networks, to user. There is multiple cloud provider, also the IaaS (Infrastructure-as-a-service) and PaaS (Platform-as-a-service). To operate multiple cloud for applications, which requires multiple private, public, or hybrid clouds, limit the capability of each cloud solution. Solution provided by such cloud will gets incompatible with others. So, to providing the solution which can compatible with multi-cloud platform is a tedious job. To achieve this CloudML provides a

"domain-specific modelling language along with run time environment" [178].

It provides the interoperability and provide vendor lock-in, also it provides the solution on specification of provisioning, deployment, and adaptation concerns of multi-cloud systems. At design time as well as runtime [178]. CloudML provides two level of abstraction while developing model for multi-cloud application: (a) Cloud Provider-Independent Model (CPIM), this specifies the provisioning and deployment. (b) Cloud Provider-Specific Model (CPSM), which filters the provisioning and deployment of multiple cloud application, according to its cloud. These two abstract approach help CloudML to achieve the multi-cloud application support [179].



title	CloudStack
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

“Apache CloudStack is an open source Infrastructure-as-a-Service platform that manages and orchestrates pools of storage, network, and computer resources to build a public or private IaaS compute cloud” [180].

In traditional IT infrastructure, resources scaling is a cumbersome task to accommodate the growing or shrinking needs of hosted applications. The cloud computing overcomes these disadvantages by using the virtualization technique to its core, one of them being the aforementioned Apache CloudStack [181].

Apache CloudStack features on-demand elastic cloud computing capabilities such as Amazon EC2. Also, it has rich management interface compatible with the latest browsers. One can use it to build public-private and hybrid cloud environments. Moreover, it supports all hypervisors for virtualization such as Hyper-V, XenServer, Xen Project, KVM-RHEL and many more. It is primarily written in Java and includes REST API for managing cloud operations and a web GUI that allows to manage and organize cloud environments to the administrators as well as to the end-users to adjust VM templates [180].

In most straightforward of the architecture, CloudStack consists of the Management server which can manage multiple geographically distributed zones and a machine that acts as a cloud infrastructure whose resources are needed to be controlled. The management server may run on the physical server or Virtual machine and

requires application server and MySQL database for persistence. The deployment architecture consists of a hypervisor (virtualization software); a cluster which is a group of identical hosts running a common hypervisor and equipped with a primary storage unique to each cluster; a pod which can have one or more clusters and L2 switches; zones that include one or more pods and access to secondary storage which consists of templates, ISO images or snapshots. A zone typically represents a single datacenter. In addition, regions represent collections of proximate zones. Regions are the largest components in the CloudStack deployment framework. Cloudstack mainly supports basic and advanced type of networking. The basic configuration resembles more to classic AWS-style networking while the advanced configuration typically uses layer-2 isolation [180].

As stated on the Apache website, The VMOps Startup started developing the CloudStack project in 2008 and changed the name to Cloud.com. Later on, in July 2011, the company was purchased and was submitted to the Apache Incubator in 2012. On March 20, 2013, Apache CloudStack graduated from the incubator, and the announcement released on March 25, 2013 [182].

Today, Apache CloudStack services used by a wide variety of users including Apple, Bell Canada, Zynga, Huawei [183]. Many of those users

“incorporate or integrate with Cloudstack in their own products, organizations who have used Cloudstack to build their own private clouds, and systems integrators that offer CloudStack related services” [183].

3.74 CNTK



title	CNTK
status	90
section	Application and Analytics
keywords	Application and Analytics

The Microsoft Cognitive Toolkit - CNTK - is a unified deep-learning toolkit by Microsoft Research. It is in essence an implementation of Computational Network (CN) which supports both CPU and GPU. CNTK supports arbitrary valid computational networks and makes building DNNs, CNNs, RNNs, LSTMS, and other complicated networks as simple as describing the operations of the networks. The toolkit is implemented with efficiency in mind. It removes duplicate computations in both forward and backward passes, uses minimal memory needed and reduces memory reallocation by reusing them. It also speeds up the model training and evaluation by doing batch computation whenever possible [184]. It can be included as a library in your Python or C++ programs, or used as a standalone machine learning tool through its own model description language (BrainScript) [185]. Latest Version:2017-02-10. V 2.0 Beta 11 Release

3.75 Cobbler



title	Cobbler
status	90
section	DevOps
keywords	DevOps

Cobbler is a Linux provisioning system that facilitates and automates the network based system installation of multiple computer operating systems from a central point using services such as DHCP, TFTP and DNS [186]. It is a nifty piece of code that assembles all the usual setup bits required for a large network installation like TFTP, DNS, PXE installation trees and automates the process. It can be configured for PXE, reinstallations and virtualized guests using Xen, KVM or VMware. Cobbler interacts with the koan program for re-installation and virtualization support. Cobbler builds the Kickstart mechanism and offers installation profiles that can be applied to one or many machines. Cobbler has features to dynamically change the information contained in a kickstart template (definition), either by passing variables called ksmeta or by using so-called snippets.

3.76 CompLearn



title	CompLearn
status	90
section	Application and Analytics
keywords	Application and Analytics

Complearn is a system that makes use of data compression methodologies for mining patterns in a large amount of data. So, it is basically a compression-based machine learning system. For identifying and learning different patterns, it provides a set of utilities which can be used in applying standard compression mechanisms. The most important characteristic of complearn is its power in mining patterns even in domains that are unrelated. It has the ability to identify and classify the language of different bodies of text [187]. This helps in reducing the work of providing background knowledge regarding a particular classification. It provides such generalization through a library that is written in ANSI C which is portable and works in many environments [187]. Complearn provides immediate access every core functionality in all the major languages as it is designed to be extensible.

3.77 CoreOS



title	CoreOS
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

[188] states that

“CoreOS is a linux operating system used for clustered deployments.”

CoreOS allows applications to run on containers. CoreOS can be run on clouds, virtual or physical servers. CoreOS allows the ability for automatic software updates inorder to make sure containers in cluster are secure and reliable. It also makes managing large cluster environments easier. CoreOS provides open source tools like CoreOS Linux, etcd,rkt and flannel. CoreOS also has commercial products Kubernetes and CoreOS stack. In CoreOS linux service discovery is achieved by etcd, applications are run on Docker and process management is achieved by fleet.



title	Couchbase Server
status	10
section	NoSQL
keywords	NoSQL

Couchbase, Inc. offers Couchbase Server (CBS) to the marketplace as a NoSQL, document-oriented database alternative to traditional relationship-oriented database management systems as well as other NoSQL competitors. The basic storage unit, a document, is a

“data structure defined as a collection of named fields”.

The document utilizes JSON, thereby allowing each document to have its own individual schema [189].

CBS combines the in-memory capabilities of Membase with CouchDB’s inherent data store reliability and data persistency. Membase functions in RAM only, providing the highest-possible speed capabilities to end users. However, Membase’s in-ram existence limits the amount of data it can use. More importantly, it provides no mechanism for data recovery if the server crashes. Combining Membase with CouchDB provides a persistent data source, mitigating the disadvantages of either product. In addition, CouchDB + membase allows the data size

“to grow beyond the size of RAM” [190].

CBS is written in Erlang/OTP, but generally shortened to just Erlang. In actuality, it is written in

“Erlang using components of OTP alongside some C/C++”.

It runs on an Erlang virtual machine known as BEAM [191] [192].

Out-of-the-box benefits of Erlang/OTP include dynamic type setting, pattern matching and, most importantly, actor-model concurrency. As a result, Erlang code virtually eliminates the possibility of inadvertent deadlock scenarios. In addition, Erlang/OTP processes are lightweight, spawning new processes does not consume many resources and message passing between processes is fast since they run in the same memory space. Finally, OTP's process supervision tree makes Erlang/OTP extremely fault-tolerant. Error handling is indistinguishable from a process startup, easing testing and bug detection [193].

CouchDB's design adds another layer of reliability to CBS. CouchDB operates in append-only mode, so it adds user changes to the tail of database. This setup resists data corruption while taking a snapshot, even if the server continues to run during the procedure [194].

Finally, CB uses the Apache 2.0 License, one of several open-source license alternatives [195].



title	CouchDB
status	10
section	NoSQL
keywords	NoSQL

3.79.1 Old

The Apache Software Foundation makes CouchDB ??? available as an option for those seeking an open-source, NoSQL, document-oriented database. CouchDB, or cluster of unreliable commodity hardware database, stores data as a JSON-formatted document [196]. Documents can consist of a variety of field types, e.g., text, booleans or lists, as well as metadata used by the software. CouchDB does not limit the number of fields per document, and it does not require any two documents to consist of matching or even similar fields [197]. That is, the document has structure, but the structure can vary by document. CouchDB coordinates cluster activities using the master-master mode by default, which means it does not have any one in charge of the cluster. However, a cluster can be set up to write all data to single node, which is then replicated across the cluster. Either way, the system can only offer eventual consistency. CouchDB serves as the basis of Couchbase, Inc's Couchbase Server [198].

3.79.2 New

Apache CouchDB ??? is an open source database software. Apache emphasizes that CouchDB

“completely embraces the web” [199].

CouchDB can take JSON documentas input. JSON stands for JavaScript Object Notation. This is a language that is simple for humans to read

and format and serves to format data. Furthermore,

"it is used primarily to transmit data between a server and a web application"[200].

Therefore, developers can use JSON documents to transmit data between servers like CouchDB and the applications they are developing [200]. CouchDB is a powerful system, not just a basic database software, because of the many features it comes with. For example, it has real-time change notifications that are characteristic of blockchain computing for accounting firms. Apache's CouchDB was built using Erlang's OTP platform which was designed with the intent of serving real-time applications.

CouchDB's signature is relax. This is because the operations of this system are simple and do not risk behavior that is unforeseeable or bugs that cannot be traced to their origin [201]. This is especially useful to developers who do not have a very strong background in database structures and operations.

Apache CouchDB is also set apart from earlier database softwares specifically because of its real time capabilities. Before CouchDB, one had to make a request by > "talking to the server, wait for the server to process the request, wait for > the result to come back, display the result. Every. Time." [202].

However CouchDB was special because it allowed the database to be accessed from a local machine instead of reaching the server every time. This increased the speed, but was limited partially by the hardware capabilities. Furthermore, CouchDb's replication features allows users to access the same data across the globe while reducing the latency that is generally accompanied by more primitive database software like MySQL. ## Crunch f18-523-53



title	Crunch
status	10
section	Workflow-Orchestration

keywords Workflow-Orchestration

Arvados Crunch is a containerized workflow engine for running complex, multi-part pipelines or workflows in a way that is flexible, scalable, and supports versioning, reproducibility, and provenance [203]. Crunch runs in virtualized computing environments. Crunch is the name for the Arvados system for managing computation. It provides an abstract API to various clouds and resource allocation and scheduling systems, and integrates closely with Keep storage and the Arvados permission system. Crunch is designed to process large volumes of data by running tasks parallelly and asynchronously. Crunch is capable of using multiple processors at once. Crunch helps in debugging by tracking inputs and outputs based on the settings the user provides. Crunch helps in running the pipelines with different versions of the code in the repository (such as GIT).

Since crunch creates workflows that run concurrently, each instance of concurrent-hash creates a separate checksum file as output. crunch automatically collates these files into a single collection, which is the output of the job. Crunch integrates with Keep and git repositories to maintain provenance. It can be used off-the-shelf software tools in distributed computations and it is efficient over wide range of problem sizes. Crunch is flexible with any programming language and execution environment. Crunch helps in isolating workloads by running all the jobs in docker containers.

Here are few key features of Arvados Crunch over Hadoop Mapreduce [204] Provenance and Reproducibility - Like Keep, the Arvados distributed file system, Crunch is designed to automatically track the origin of result data. It can also efficiently reproduce complex workflows and comparing workflows to one another.

Performance - Most genomics problems are embarrassingly parallel and can benefit from horizontal scaling. In the cloud, Crunch can deliver cost-effective performance for genomics related analyses by automatically adjusting the available compute resources to the workload.

Standardization - Common Workflow Language is the workflow description standard in bioinformatics. It is the native workflow language in Crunch.

3.80 CUBRID



title	CUBRID
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

CUBRID name is deduced from the combination of word CUBE (security within box) and BRIDGE (data bridge). It is an open source Relational DataBase Management System designed in C programming language with high performance, scalability and availability features. During its development by NCL, korean IT service provider the goal was to optimize database performance for web-applications [205]. Importantly most of the SQL syntax from MYSQL and ORACLE can work on cubrid.CUBRID also provides manager tool for database administration and migration tool for migrating the data from DBMS to CUBRID bridging the dbs. CUBRID enterprise version and all the tools are free and suitable database candidate for web-application development.



title	CUDA
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Compute Unified Device Architecture (CUDA) is a parallel computing API, [206] and a general-purpose GPU (graphics processing unit) developed by NVIDIA where the code written, interacts directly with the GPU. It is used as the computing engine and has a unique architecture well suited to tackle large computations in big data processing. CUDA, which was used to render video and images, are also being used to solve mathematics problems which are computationally intensive in a cost-effective way. CUDA based systems are used in the training of deep learning algorithms [206].

CUDA based systems has drastically reduced the time taken to train those algorithms that process large data sets in parallel to just few hours compared to the CPU based systems, which usually takes a long time. The CUDA GPUs are also used in a wide variety of applications that are not related only to graphics:

- They are used in the field of computer vision and speech recognition by leveraging deep learning.
- Data mining and analytics, Medical imaging, Geo-intelligence etc [207].

CUDA shares some of its computing interface with Open Computing Language (OpenCL) even though it has its own application programming Interface (API) [208].

"The CUDA Architecture included a unified shader pipeline, allowing each and every arithmetic logic unit

(ALU) on the chip to be marshaled by a program intending to perform general-purpose computations" [206].

CUDA supports most Windows, Linux, and Mac OS compilers. Hadoop platform using Nvidia CUDA architecture was used as a solution to handle fastest growing data in form of platforms or infrastructures for processing/computation [206].

The CUDA processing flow has four main steps:

Copy data from main memory to GPU memory, CPU instructs the process to GPU, GPU executes parallel in each core, Copy the results from GPU to main memory. [209]

Some examples where CUDA is used:

Video file format interconversion, 3D Graphics generation, Compression of files Face recognition, Distributed Computing [209].

Some more examples include Molecular dynamics, Simulating the motion of fluids using the numerical methods, Environmental Science [208].

3.82 D3.js



title	D3.js
status	90
section	Application and Analytics
keywords	Application and Analytics

D3.js is a JavaScript library responsible for manipulating documents based on data. D3 helps in making data more interactive using HTML, SVG, and CSS. D3's emphasis on web standards makes it framework independent utilizing the full capabilities of modern browsers, combining powerful visualization components and a data-driven approach to DOM manipulation [210].

It assists in binding random data to a Document Object Model (DOM), followed by applying data-driven transformations to the document. It is very fast, supports large datasets and dynamic behaviours involving interaction and animation.



title	DAAL (Intel)
status	10
section	Application and Analytics
keywords	Application and Analytics

3.83.1 Old text

DAAL stands for Data Analytics Acceleration Library. DAAL is software library offered by Intel which is written in C++, python, and Java which implements algorithm for doing efficient and optimized data analysis tasks to solve big-data problems [211]. The library is designed to use data platforms like Hadoop, Spark, R, and Matlab. The important algorithms which DAAL implements are 'Lower Order Moments' which is used to find out max, min standard deviation of a dataset, 'Clustering' which is used to do unsupervised learning by grouping data into unlabelled group. It also include 10-12 other important algorithms.

DAAL supports three processing modes namely batch processing, online processing and distributed processing [212]. Intel DAAL addresses all stages of data analytics pipeline namely pre-processing, transformation, analysis, modelling, validation, and decision making.

3.83.2 New text

try to start with a small sentence about Daal that summarizes what it is and you wrote that is not quoted

improve grammar

The Intel Data Analytics Acceleration Library is design to accerler the data analysis process by providing integrated functions.

"The Intel Data Analytics Acceleration Library (Intel DAAL) helps speed big data analytics by providing highly optimized algorithmic building blocks for all data analysis stages (Pre-processing, Transformation, Analysis, Modeling, Validation, and Decision Making) for offline, streaming and distributed analytics usages. It is designed for use with popular data platforms including Hadoop, Spark, R, and Matlab. for highly efficient data access" [213].

To make the library being widely used, Intel offers the different interfaces for different programming languages such as Python, Java and C++, which allows developers with different skill background have access to the new technology.

There are several components in the data analytics acceleration library which supports the acceleration of the data analysis process. The first one is Data Management component is a module of classes and utilities for combining the dataset, pre-processing data and make the data to be universal to different platforms, and process the formats of the data for the follow-up steps. The second part is the core part, Algorithms, which includes common algorithms for data analysis, machine learning, and model training. Besides, the data analytics acceleration library also involves the Services part for the link between the previous two parts.

The Intel data analytics acceleration library's Algorithms contains lots of useful algorithms ranging from basic data mining algorithms to higher level machine learning algorithms. The low order moments could calculate the basic features for the data such as min, max, mean, standard deviation, etc. And also some other features which could be useful to other kinds of analysis such as ANOVA table. The quantiles could show the different groups distribution clearly, which is very common in the research fields. The correlation matrix and variance-covariance matrix could help us have a basic understanding of a dataset, which includes the tendency and dependence among variables. The regression part is a simple to find the relationship

between two datasets, such as the simplest one - the Titanic dataset from Kaggle.

3.84 Databus



title	Databus
status	10
section	Streams
keywords	Streams



title	DataFu
status	10
section	Application and Analytics
keywords	Application and Analytics

3.86 Old

The Apache DataFu project was created out of the need for stable, well-tested libraries for large scale data processing in Hadoop. Apache DatFu consists of two libraries Apache DataFu Pig and Apache DataFu Hourglass [214]. Apache DataFu Pig is a collection of useful user-defined functions for data analysis in Apache Pig. The functions are in areas of Statistics, Bag Operations, Set Operations, Sessions, Sampling, Estimation, Hashing and Link Analysis. Apache DataFu Hourglass is a library for incrementally processing data using Hadoop MapReduce. It is designed to make computations over sliding windows more efficient. For these types of computations, the input data is partitioned in some way, usually according to time, and the range of input data to process is adjusted as new data arrives. Hourglass works with input data that is partitioned by day, as this is a common scheme for partitioning temporal data.

3.87 New

DataFu [214] is a collection of libraries developed by Apache that aid in the use of data mining and statistical methods within big data environments such as Hadoop. Datafu consists of two main libraries; Pig, which is a collection of user defined functions, and Hourglass, a processing framework within a tool called MapReduce. The Pig library began and was open sourced in 2010. During the last several years, the application has continued to received numerous contributions. The Hourglass library concept was presented at a IEEE Big Data conference in 2013, and hence began receiving contributions and is in widespread usage at large organizations such as LinkedIn [215].

DataFu's Pig application contains a wide array of libraries that assist users in working with very large datasets. Pig includes a standard statistics library which includes functions to compute mean, median, quantiles, confidence intervals and others. Pig includes functions that pertain to set operations such as finding set intersections and unions. Pig incorporates functions that work with data bags.

"A data bag is a global variable that is stored as JSON data and is accessible from a Chef server. A data bag is indexed for searching and can be loaded by a recipe or accessed during a search. The contents of a data bag can vary, but they often include sensitive information (such as database passwords)" [216].

DataFu Pig can also perform tasks pertaining to data sampling, estimation, link analysis, and data sessionizing. DataFu Hourglass was designed to work with very large computations over a sliding window timeframes more effectively, using partitioned data over some time frame.

"Hourglass works with input data that is partitioned by day, as this is a common scheme for partitioning temporal data" [215].

Hourglass was designed with the following two computational models in mind; fixed length vs fixed-start:

"Fixed-length: the length of the window is set to some constant number of days and the entire window moves forward as new data becomes available. Example: a daily report summarizing the the number of visitors to a site from the past 30 days. Fixed-start: the beginning of the window stays constant, but the end slides forward as new input data becomes available. Example: a daily report summarizing all visitors to a site since the site launched" [215].



title	DataNucleus
status	10
section	Object-relational mapping
keywords	Object-relational mapping

DataNucleus [217] is an advanced form of the Java Persistent Objects (JPOX) technology. It is an open-source programming software product used for data management in Java[218]. It supports a wide range of datastores and provides persistence to relational, graph-based, map-based, object-based datastores, document-based, and web-based storage. The software is mainly used to persist classed data to the datastore. Persistence is the property of outliving the process that created the system. For instance, after a computer breakdown, the hard-disk remains stored with information that can still be retrieved and used in another computer instead of losing all the info with other programs. The DataNucleus persists the stored values and sets them again when retrieving data from the datastore. This function is made possible by the use of standard Application Programming Interfaces (APIs). Java has two APIs: Java Data Objects (JDO) and Java Persistence API (JPA). The difference between the two is mainly the metadata definitions applied.

The DataNucleus technology is applied by system developers to optimize the user experience of their software. It gives the user control of what he wants to retrieve through a customizable interface. The DataNucleus also enables interface interoperability for users who wish to change their datastores in the future. It has an extension mechanism for its interfaces which can be extended to allow it to perform more functions. An interactive user interface makes the computer programmer comfortable with the system because it can be customized to suit his preferences or those of his client.

The technology also helps when a computer breaks down to save and retrieve valuable information by preventing data losses. The process is very quick and easy. And it provides outstanding performance when compared with the competing technologies.



title	DataTurbine
status	10
section	Streams
keywords	Streams

Data turbine [219] is an open source project which enable users to retrieve information and data live from different data sources. Information or data are not limited to commonly known data sources, i.e. videos, images, or texts, but also in various formats from different devices, i.e. web cam videos, experiment or lab results. The usage of data turbine are more emphasized on streaming data which accelerated by the systems to efficiently display information in faster rate as compare to regular publish or subscribe system. Processing live stream data as a middleware, data turbine enables users to interact with the information flexibly; users are not only restricted to view the stream in one direction, but also allows them to rewind or pause the stream [220].

Data Turbine utilizes network bus objects and ring buffers. Network bus objects work in such a way that it combines multiple data streams and streamlines into uniform framework. As such, it can process different data sources from variety devices by uniformizing the data that can be integrated into different API. The ring buffers, on the other hand, works as “tunable persistent storage at key network nodes” which allows the flexibility to stream the data ,i.e. rewinding, replaying and/or capturing. With both of these combined, data turbine are most commonly used for sensor-based systems where data are source constantly send and receive data [221].

Users that use data turbine are most commonly working in the field of science and engineering, ranging from ecology to aerospace. The Coral Reef Environmental Observatory Network (CREON) are one of

the example of an organization that utilize the platform to monitor marine habitats. Working with sensors and monitoring the data through live streaming, data turbine allows the organization to monitor these sensors in a huge scale and from different data center. As such, sensors from different regions can also be integrated into the systems [221].

3.90 DB2



title	DB2
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

DB2 is a Relational DataBase Management System (RDBMS). Though initially introduced in 1983 by IBM to run exclusively on its MVS (Multiple Virtual Storage) mainframe platform, it was later extended to other operating systems like UNIX, Windows and Linux. It is used to store, analyze and retrieve the data and is extended with the support of Object-Oriented features and non-relational structures with XML [222]. DB2 server editions include: Advanced Enterprise Server Edition and Enterprise Server Edition (AESE / ESE) designed for mid-size to large-size business organizations, Workgroup Server Edition (WSE) designed for Workgroup or mid-size business organizations, Express -C provides the capabilities of DB2 at no charge and can run on any physical or virtual systems, Express Edition designed for entry level and mid-size business organizations, Enterprise Developer Edition offers single application developer useful to design, build and prototype the applications for deployment on the IBM server. DB2 has APIs for REXX, PL/I, COBOL, RPG, FORTRAN, C++, C, Delphi, .NET CLI, Java, Python, Perl, PHP, Ruby, and many other programming languages. DB2 also supports integration into the Eclipse and Visual Studio integrated development environments [223].



title	DC.js
status	10
section	Application and Analytics
keywords	Application and Analytics

DC.js [224] is an open-source framework written entirely in JavaScript to visualize data on the web in a web-browser. Data Visualization is an important process that is imperative before a decision-making process in any business these days. This has led to the rise of various tools and frameworks for developers to analyze enterprise level complex real-world data, DC.js is one such popular framework, since it is derived from the popular D3.js (Data-Driven-Documents) framework it supports exploration on large multi-dimensional complex real-world datasets. D3 [210] has been the de-facto standard to build rich, interactive visualizations on the web. DC.js uses d3 framework to render charts in a web browser friendly format. The charts rendered using JavaScript framework on the client-side browser are data driven and reactive and therefore, provide instant feedback to user interaction and are easy to interact with.

The DC.js framework can be used to perform exploratory data analysis across platforms in various browsers. The developers or users can choose from a plethora of charts ranging from the simple boxplot to a complex heatmap or a map having geographical coordinates to view their data. All of these can be developed using any latest Integrated Development Environments (IDEs) such as Visual Studio, Eclipse. To get started with charts in DC.js the developer or user need not have any prior knowledge of D3.js [225]. DC.js is mainly built using two libraries in JavaScript, those are:

- Crossfilter: This library is primarily used for exploring datasets that contain millions of records with many dimensions or

attributes and enables developers to create super-fast interactive visualizations. It is also used to perform aggregation operations on millions of rows at a rapid speed.

- D3.js: D3.js framework is primarily used to build rich, interactive visualizations on the web. It was created by Mike Bostock and has since been used by all major websites to create rich and interactive visualizations, charts and dashboard on websites all over the world.

3.92 DevOpSlang



title	DevOpSlang
status	90
section	DevOps
keywords	DevOps

DevOpSlang serves as means of collaboration and provides the foundation to automate deployment and operations of an application. Technically, it is a domain specific language based on JavaScript Object Notation (JSON). JSON Schema is used to define a formal schema for DevOpSlang and complete JSON Schema definition of DevOpSlang is publicly available on GitHub project DevOpSlang: <http://github.com/jojow/devopslang> Devopsfiles are the technical artifacts (Unix shell commands, Chef Scripts, etc.) rendered using DevOpSlang to implement operations. Beside some meta data such as 'version' and 'author' Devopsfile defines operations like 'start' consisting of a single or multiple actions which specifies the command to run the application. Similarly, a 'build' operation can be defined to install the dependencies required to run the application. Different abstraction levels may be combined consistently such as a 'deploy' operation consisting of actions on the level of Unix shell commands and actions using portable Chef cookbooks [226].



title	Disco
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

3.93.1 Old Text

Disco [227] is an implementation of mapreduce for distributed computing that benefits end users by relieving them of the need to handle

“difficult technicalities related to distribution such as communication protocols, load balancing, locking, job scheduling, and fault tolerance” [228].

Its designers wrote the software in Erlang, an inherently fault tolerant language. In addition, Disco’s creators chose Erlang because they believe it best meets the software’s need to handle

“tens of thousands of tasks in parallel” [229].

Python was used for Disco’s libraries. Finally, Disco supports pipelines,

“a linear sequence of stages, where the outputs of each stage are grouped into the input of the subsequent stage” [230].

Its designers implemented Disco’s libraries in Python. Disco originated within Nokia Corp. to handle large data sets. Since then it

has proven itself reliable in production environments outside of Nokia [231].

- b. DISCO from the research group Service Engineering (SE), serves as “an abstraction layer for OpenStack’s orchestration component [Heat]” SE based DISCO on its prior orchestration framework, Hurtle. The software sets up a computer cluster and deploys the user’s choice of distributed computing architecture onto the cluster based on setup inputs provided by the user [228]. DISCO offers a command line interface via HTTP to directly access OpenStack [228].

3.93.2 New Text

Disco [227] takes on the lofty task of attempting to be a substitute for a Hadoop Distributed File System in a light weight, Python implemented method for polling data. While Hadoop is the industry leader in distributed filesystems, Disco offers simplified coding, concepts, and implementation, appealing to a wide range of data users. The libraries involved are still in the works but deployment is very simple, especially in a Python-centric organization/SAN environment [227]. The access is simplified and diverse as it can leverage many different access protocols. APIs exist to simplify. It otherwise is similar to Hadoop’s MapReduce but instead of consisting of <key,value> pairs it follows more a database infrastructure. Disco uses an opensource distribution that can be easily found on GitHub, as mentioned in the references.

“Disco is a distributed map-reduce and big-data framework. Like the original framework, which was publicized by Google, Disco supports parallel computations over large data sets on an unreliable cluster of computers” [227].

Big data analytics is dramatically shifting over to Python, despite lack of computing efficiency, because of its robust libraries and simplified programming styles [227]. From this, it was only a matter of time until

the data storage and distribution took on a Python implementation for a distributed file system. The intention is that the analytics and storage will still be able to be offloaded to other compute powers instead of simply on one local machine [227]. The benefit of ease for programmers can now be enjoyed, potentially without needing to learn the intricacies of SQL/NoSQL and other, more complicated, programming languages.

The Disco project is relatively new and young but has high promise, especially in Python implemented environments. The concept of a distributed file system is that, in big data analytics, many times the data is too large to be both stored and analyzed on a single machine. While super-computers are one way around needing such a program, cheaper implementations, using existing hardware and smaller storage clusters is viable and valuable.

3.94 Distributed Coordination



title	Distributed Coordination
status	90
section	Monitoring
keywords	Monitoring

3.95 DL4j



title	DL4j
status	90
section	Application and Analytics
keywords	Application and Analytics

DL4j stands for Deeplearning4j [232]. It is a deep learning programming library written for Java and the Java virtual machine (JVM) and a computing framework with wide support for deep learning algorithms. Deeplearning4j includes implementations of the restricted Boltzmann machine, deep belief net, deep autoencoder, stacked denoising autoencoder and recursive neural tensor network, word2vec, doc2vec, and GloVe. These algorithms all include distributed parallel versions that integrate with Apache Hadoop and Spark. It is a open-source software released under Apache License 2.0.

Training with Deeplearning4j occurs in a cluster. Neural nets are trained in parallel via iterative reduce, which works on Hadoop-YARN and on Spark. Deeplearning4j also integrates with CUDA kernels to conduct pure GPU operations, and works with distributed GPUs.



title	Docker Compose
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Docker Compose is a technology that allows one to run multiple applications/services at the same time while using a mutual host [233]. It helps development teams and administrators that are in need of a framework to be able to handle rapid changes and scaling out as part of the DevOps process. The aforementioned host is usually a virtual network, which allows an increase of security and performance. The main idea behind Docker is to provide a

"lightweight environment where code can be run efficiently" [234]

as well as adding an

"extra facility to the proficient work process" [234].

The applications within Docker are packaged into individual containers which are then run as isolated processes [235]. Docker containers are created using Docker Images which are also placed into Docker registries [234]. The registries can be public (Docker Hub) or private [234]. Another important aspect of this technology is the Docker Engine which is a component added to the host operating system, more specifically as a

"layer in between the host operating system and where the applications are executed" [234].

This is the main difference between the Docker technology and VM's, as VM's require an extra layer between the host OS and guest OS

known as Hypervisor [234]. One of the benefits of the Docker technology is that

“software tools shipped via Docker remain usable for a wide audience even when the underlying dependencies evolve dramatically or when active development has ceased” [235].

Even though this technology answers the high demands of today's development process where cost reduction, short deadlines, and beating a competitor are main goals, Docker containers can manifest some drawbacks. Some of them include security concerns, inability to provide the complete virtualized environment as it depends on the Linux kernel, difficulty of running on older machines, as well as the fact that it only supports 64-bit machines [234].

3.97 Docker (Machine, Swarm)



title	Docker (Machine, Swarm)
status	90
section	DevOps
keywords	DevOps

Docker is an open-source container-based technology. A container allows a developer to package up an application and all its parts including the stack it runs on, dependencies it is associated with and everything the application requires to run within an isolated environment. Docker separates Application from the underlying Operating System in a similar way as Virtual Machines separates the Operating System from the underlying hardware. Dockerizing an application is lightweight in comparison with running the application on the Virtual Machine as all the containers share the same underlying kernel, the Host OS should be same as the container OS (eliminating guest OS) and an average machine cannot have more than few VMs running on them.

Docker Machine is a tool that lets you install Docker Engine on virtual hosts, and manage the hosts with docker-machine commands [236]. You can use Machine to create Docker hosts on your local Mac or Windows machine, on your company network, in your data center, or on cloud providers like AWS or Digital Ocean. For Docker 1.12 or higher swarm mode is integrated with the Docker Engine, but on the older versions with Machine's swarm option, user can configure a swarm cluster. Docker Swarm provides native clustering capabilities to turn a group of Docker engines into a single, virtual Docker Engine.

“With these pooled resources user can scale out your application as if it were running on a single, huge computer” [237].

Docker Swarm can be scaled up to 1000 Nodes or up to 50,000 containers



title	Dokku
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

According to its documentation, Dokku is

"The smallest Pass implementation you have ever seen" [238].

It is an open source platform which helps in managing several applications in a single server of your choice. One can deploy their application with a single command line input to cloud with as little an infrastructure cost as possible. The ideal of Dokku is to provide a platform which takes in developer's code from their laptops or systems into the cloud as efficiently as possible. That is, it takes care of the deployment process, leaving the developer to have one less thing to worry about.

The documentation in the home website of Dokku provides quick start information to set up your own version of Dokku [239]. Dokku you need atleast 1GB of the system memory to run efficiently [240]. Additionally, you can get Dokku installed on virtual machine as well and it can also be customized to suit the developers needs during the installation process. The getting started documentation by Dokku team has very clear steps to start with installation of Dokku to deployment of your application, making it very easy to use [240].

Dokku comprises of several scripts which utilizes modern bash, as bash is relatively common and has a high ease of use. These scripts work together as a pipeline which takes the code from the developer and deploys it into a successful application. The developer does not

need to worry about configuring his application or database servers. The moment the user enters

```
git push dokku master
```

the entire code is taken from the github repository from the developer and converted and deployed as an application.

The application deployed can be managed by a set of official plugins, which allow the developer to manage the environment variables, to check if the application has started properly and a storage plugin which confirms the design requirements. According to the original documentation, Dokku requires less downtime to change containers when the new code is pushed, leaving a seamless transition from code to application. Dokku, aims at taking less downtime for application deployment, giving the coders more time to work.

3.99 dotCloud



title	dotCloud
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

dotCloud services were shutdown on February 29,2016 [241].



title	Dream:Lab
status	10
section	Application and Analytics
keywords	Application and Analytics

DreamLab is app only based service developed and propagated by Vodafone Australia that helps to solve cancer problems. Since cancer has become a major problem in the world and billions of dollars are spent every year for medical treatment and diagnosis, it is a very serious issue that needs to be addressed in various possible ways, especially using technology advancements and DreamLab is one of those many ways, using mobile as platform. The researchers analyze genomic patterns like genetic mutations of cancer patients and cluster them based on their pattern. DreamLab helps to support major cancers like breast, prostate in the initial state. Later, the project is expected to be developed to support all 24 types of cancers. The genomic sequencing is a task that involves huge giga bytes of data and requires huge computational power like GPUs [242]. Since the access to such resources is very limited and involves high cost, Vodafone has developed DreamLab that enables users to donate processing power of their smartphones, which can be used for cancer research on hands. This is similar to distributed computing on cloud [243]. It uses Network connectivity analyzer algorithm, developed by the Gravan Institute of Medical Research. The algorithm calculates various statistics of genomic interactions between different sets of genes. This can help find patterns in the groups or individuals who share similar genomic mutations by using big data. The app requests research job on phone to AWS Amazon Cognito and the job is handled by AWS DynamoDB. This way, the computational power of smartphones of several users can be used for performing the algorithm process to solve the cancer research problem. Once the computations are done and any similarities in genomic mutations or

patterns are detected, the results are sent from app to Graven research team, where they use the data and findings for analysis and research. As per the Graven research study, 33 android devices would be able to process the task at as much speed as Graven super computer and also with over 50000 plus active app users and the total computational power available to them is almost 900 times that of Graven Research super computer [244].



title	Drill
status	10
section	High level Programming
keywords	High level Programming

Drill, according to its documentation,

“is an open-source software framework that supports big data exploration” [245].

Using Drill, we can join data from multiple data stores using just one single query. Drill supports collaborative exploration of large-scale datasets. Drill is the open source version of Google’s Dremel framework which can be accessed through Google BigQuery. An incomparable property of Drill is that it can be scaled to a large number of servers and has the ability to process huge amounts of data in a very short period. HBase, MongoDB, MapR-DB, HDFS, MapR-FS, Amazon S3 are few of the NoSQL databases and file systems that Drill supports.

For large-scale data processing, Drill employs a distributed system. It provides a service that is responsible for receiving requests, processing the requests and delivering the results to the clients. This service is an essential entity of Drill and is known as Drillbit. The main components of the Drillbit server are an SQL Parser, storage plug-ins, an RPC end-point, an optimizer and an execution block. It also contains a cache and a storage engine interface. All drillbits are essentially the same which is why we need another service to maintain the clusters, the most compatible one is zookeeper. Zookeeper is used to handle the memberships and also do regular check on the drillbits [246].

The flow with in drillbits are very efficient. When a client posts a query, any drillbit from the server can accept the query and instantly become a driving-drillbit. The driving-drillbit then gets a quote from the zookeeper about the availability of the rest of drillbits. It then allocates multiple sub-queries to the appropriate drillbits for execution. After each of the drillbits have finished their tasks, the results are then returned to the client [247].

The main advantage of Drill is that it has the ability to run SQL-queries at very fast rates without any delays due to its schema-free nature. Its other features include dynamic querying, reprieve from ETL, support for nested data, its ability to integrate with Hive. It has very good connectivity to multiple data bases due to its pluggable configuration.



title	Dryad
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration, retired

Dryad is a retired technology.

Dryad [248] was a project that was developed and maintained by the Microsoft Research Team in 2006. This was their solution to work with data applications parallelly in a distributed system environment. Dryad mainly deals with distributed execution graphs. Dryad focuses on the throughput instead of latency and does not focus much on security since it assumes that a private data center is used. The exponential increase of data, structured and unstructured, led to need to such infrastructure that could use a cluster of computers mostly Microsoft Windows servers instead of a single machine for processing data parallelly. Dryad made it possible for the developers to utilize a group of machines inter-connected to run algorithms that deal with large volumes of data in a parallel manner without the need for an understanding of the concept of concurrent programming. Dryad's library was implemented primarily using C++. When data is given as input to a Dryad, a Dryad job creates a graph with directed edges, but this graph is not cyclic, it is known as a Directed Acyclic Graph (DAG) [249]. In the directed acyclic graph created by Dryad, each vertex is a program and the edges represent data channels. This generated graph from the Dryad job was dynamic and could change amidst execution. The data flow from one vertex to another is realized by TCP/IP streams, shared memory, or temporary files. Dryad automatically handles the following task, manage scheduling, distribution, fault tolerance and job creation. This made Dryad perform better compared to its main competitor MapReduce which was developed by Google for the same purpose of handling large

volumes of unstructured data. This made Dryad very handy and led to the development of an entire new ecosystem to be built on top of Dryad using tools such as SSIS and DryadLINQ. But when Hadoop was open-sourced by Google, it was seen as the cheaper and viable option for developers when compared to Dryad. This made Microsoft discontinue their efforts of integrating Dryad into their Azure Cloud Platform. In 2011, Microsoft announced it was focusing their efforts on Hadoop instead of Dryad [250].



title	e-Science Central
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

e-Science Central (e-SC) [:o ref missing] is an open source cloud-based data analysis platform. It provides software as a service (SaaS) for storage and data analysis for scientists. E-SC can be deployed public clouds like Amazon AWS, Microsoft Azure or private clouds. Scientists can upload their data to the cloud using web interface. Scientists can also share data with other scientists by giving access permission or make the data public. Public data can be accessed through web search services. e-SC also supports creating groups, group members can access the data and the code; this way scientists can collaborate easily. When a change is made on data, it versions the data to allow reproduce the experiment later time.

E-science Central provides API to

“allow users to develop and upload new services to run on the cloud platform and for external applications to access data, code and workflows deployed within e-SC.” [251]

Scientists can analyze their data creating workflows using Workflows editor. Thanks to editor, users can create workflows just dragging services and connect them. Users can also create their own customized services in Java, R, Octave, C#, and JavaScript.

“A core set of services are provided for data manipulation, statistic analysis and visualization. However, the e-Science Central Science Platform as a

Service allows developers can upload their own services into the system and share them in a controlled way, as for data." ???

After running a workflow, users can see the results in a web browser and store them. If an application has multiple workflows, e-SC deploys services on multiple machines and executes them concurrently. Thus, the calculation can be made much faster.

Workflow services have data input, output ports and these ports restrict the data types. e-SC workflow engine only supports three data types:

`data-wrapper`: rectangular data that have rows and columns. Each row represents an instance.

`file-wrapper`: a file or list of files. Since the workflow system doesn't know the content of the file, interpretation must be handled by service code.

`object-wrapper`: a serialized Java object [251].

3.104 EclipseLink



title	EclipseLink
status	10
section	Object-relational mapping
keywords	Object-relational mapping

EclipseLink is an open source persistence Services project from Eclipse foundation. It is a framework which provide developers to interact with data services including database and web services, Object XML mapping etc [252]. This is the project which was developed out of Oracle's Toplink product. The main difference is EclipseLink does not have some key enterprise feature. Eclipselink support a number of persistence standard model like JPA, JAXB, JCA and Service Data Object. Like Toplink, the ORM (Object relational model) is the technique to convert incompatible type system in Object Oriented programming language. It is a framework for storing java object into relational database.

3.105 Eduroam



title	Eduroam
status	90
section	Monitoring
keywords	Monitoring

Eduroam is an initiative started in the year 2003 when the number of personal computers within the academia are growing rapidly. The goal is to solve the problem of secure access to WI-FI due to increasing number of students and research teams becoming mobile which was increasing the administrative problems for provide access to WI-FI. Eduroam provides any user from an eduroam participating site to get network access at any institution connected through eduroam. According to the organization it uses a combination of radius-based infrastructure with 802.1X standard technology to provide roaming access across research and educational networks. The role of the RADIUS hierarchy is to forward user credentials to the users home institution where they can be verified. This proved to be a successful solution when compared to other traditional ways like using MAC-address, SSID, WEP, 802.1x (EAP-TLS, EAP-TTLS), VPN Clients, Mobile-IP etc which have their own short comings when used for this purpose [253]. Today by enabling eduroam users get access to internet across 70 countries and tens of thousands of access points worldwide [254].

3.106 Ehcache



title	Ehcache
status	90
section	In-memory databases/caches
keywords	In-memory databases/caches

EHCACHE is an open-source Java-based cache. It supports distributed caching and could scale to hundred of caches. It comes with REST APIs and could be integrated with popular frameworks like Hibernate [255]. It offers storage tires such that less frequently data could be moved to slower tires [256]. It is XA compliant and supports two-phase commit and recovery for transactions. It is developed and maintained by Terracotta and is available under Apache 2.0 license. It conforms to Java caching standard JSR 107.



title	Elasticsearch
status	10
section	Application and Analytics
keywords	Application and Analytics

Elasticsearch is a distributed RESTful search engine built for the cloud. The distributed search and analytics engine is built on Apache Lucene[257].

The data is sent in JSON format to Elasticsearch using the API or ingestion tools such as Logstash and Amazon Kinesis Firehose. Elasticsearch automatically stores the original document and adds a searchable reference to the document in the cluster index. The document can be searched using the Elasticsearch API. Kibana, an open-source visualization tool can be used with Elasticsearch to visualize data and build interactive dashboards.

“It is a open source software and can be run on Amazon EC2, or on Amazon Elasticsearch Service [258].”

Amazon Elasticsearch Service is a fully managed service, and we do not have to worry about time-consuming cluster management tasks such as hardware provisioning, software patching, failure recovery, backups, and monitoring[259].

“Elasticsearch is an open-source, broadly-distributable, readily-scalable, enterprise-grade search engine. Accessible through an extensive and elaborate API, Elasticsearch can power extremely fast searches that support your data discovery applications.[260]”

Key Features of elasticsearch are as follows[258]:

- Distributed and Highly Available Search Engine
 - Each index is fully sharded with a configurable number of shards. Each shard can have one or more replicas. Read / Search operations performed on any of the replica shards.
- Multi Tenant.
 - Support for more than one index. Index level configuration (number of shards, index storage).
- Various set of APIs
 - HTTP RESTful API
 - Native Java API. All APIs perform automatic node operation rerouting.
- Fast time-to value
 - Elasticsearch offers simple REST based APIs, a simple HTTP interface, and uses schema-free JSON documents, making it easy to get started and quickly build applications for a variety of use-cases.
- High Performance
 - The distributed nature of Elasticsearch enables it to process large volumes of data in parallel, quickly finding the best matches for queries.
- Complimentary Tooling And Plugins
 - Elasticsearch comes integrated with Kibana, a popular visualization and reporting tool. It also offers integration with Beats and Logstash, while enable you to easily transform source data and load it into your Elasticsearch cluster. A number of open-source Elasticsearch plugins can be used such as language analyzers and suggests to add rich functionality to the applications.
- Near Real-Time Operations
 - Elasticsearch operations such as reading or writing data usually take less than a second to complete. This lets us use Elasticsearch for near real-time use cases such as application monitoring and anomaly detection.
- Easy Application Development

- Elasticsearch provides support for various languages including Java, Python, PHP, JavaScript, Node.js, Ruby, and many more.

3.108 Engine Yard



title	Engine Yard
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

A deployment platform with fully managed services that combines high-end clustering resources to run Ruby and Rails applications in the cloud is offered by Engine Yard. It is designed as a platform-as-a-Service for Web application developers using Ruby on Rails, PHP and Node.js who requires the advantages of cloud computing. Amazon cloud is the platform where the Engine Yard perform its operations and accomplishes application stack for its users. Amazon allows as many as eight regions to Engine Yard to deploy its CPU instances in varying capacities such as normal, high memory and high CPU. According to customer requirements multiple software components are configured and processed when an instance is started in Engine Yard.

Engine Yard builds its version on Gentoo Linux and has non-proprietary approach to its stack. The stack includes HAProxy load balancer, Nginx and Rack Web servers, Passenger and Unicorn app servers, as well as MySQL and PostgreSQL relational databases in addition to Ruby, PHP, and Node.js. The credibility of Engine Yard rests with orchestration and management as developers have option of performing functions in Amazon cloud. Standard operations management procedures are performed once the systems are configured and deployed. Key operations tasks such as performing backups, managing snapshots, managing clusters, administering databases and load balancing are taken care by Engine Yard.

Engine Yard users are empowered as they have more control over virtual machine instances. These instances are dedicated instances

and are not shared with other users. As the instances are independent every user can exercise greater control over instances without interferences with other users [261].



title	Espresso
status	10
section	NoSQL
keywords	NoSQL

Espresso is a dedicated online distributed data platform developed by LinkedIn for its applications data storage, to replace the earlier traditional relational databases for some of reasons like schema evolution, which is pretty hard and painful in RDBMS to change the tables and suspending the services due the process, Data Center failure as the RDBMS operated on master slave mode where any failure in data center required several manual hours and resulted in temporary service shutdown and Cost, RDBMS is way expensive due to hardware and associated software installations and maintenance. RDBMS could also not offer scalability, consistency, and distributed computing for parallel processing. It has a hierarchical data model, which is similar to that of RDBMS i.e., database, table, collection, document. It has a transactional support for changes to documents, secondary indexing, schema evolution, and consistency on data capture stream. The Espresso data model was designed based on use cases and access patterns at LinkedIn and is different from key value data model, which is not scalable [262]. Espresso hierarchical data model is akin to Nested entity model efficiently. Espresso offers REST API for integration into main application pretty easily that supports the following operations:

- Read operations
- Write operations
- Conditionals
- Multi operations
- Change stream listener

In read operations, the documents are retrieved by using primary key and leading key. Espresso can support advanced operations in write like partial update to documents in the database, automated increment of partial key and transactional update to groups of documents. Conditionals supported on either of the above operations i.e., reading and writing. Simple conditionals like recently changed documents list to very complex conditionals on multiple attributes. Multi operations can perform batch operation based on multiple operations like read, write conditionals etc. grouped into one. Espresso API also provides another service called Change Streamer Listener that allows us to observe any changes happening in the database. This can be used to monitor and prevent external forces in making any changes to the database. All these features allow Espresso to handle bulk and fast data flow efficiently and consistently without any failures [263].

3.110 Eucalyptus



title	Eucalyptus
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Eucalyptus is a Linux-based open source software framework for cloud computing that implements Infrastructure as a Service (IaaS). IaaS are systems that give users the ability to run and control entire virtual machine instances deployed across a variety physical resources [264]. Eucalyptus is an acronym for

“Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems.”

A Eucalyptus private cloud is deployed on an enterprise’s data center infrastructure and is accessed by users over the enterprise’s intranet. Sensitive data remains entirely secure from external interference behind the enterprise firewall [265].



title	Event Hubs
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Event Hubs [266] is a streaming platform for Big Data, which can process millions of information called events in any second. It can receive, process and store data produced by several distributed services, and can be transformed using any existing analytics tool. Event Hubs is ideally used for Anomaly detection, archiving data, live reporting, analytics pipelines, application logging and so on. It helps in easier data processing and analytics to gain timely knowledge from the real time data from various sources [267].

Event Hubs provides an event big data pipeline which ensures a proper communication platform between data or event publishers and their respective consumers. It also allows a stream handling with different characteristics from the traditional ones. Each of the capabilities of Event Hubs is built around a very huge processing scenario. Event Hubs supports various protocols for publishing data events into Kafka ecosystems by Apache [268]. Its key features include the fully managed configuration of PaaS service, it supports real-time processing, its highly scalable, and has key components in architecture which are discussed in brief below.

Features:

1. Fully Managed Configuration of PaaS service:

Event Hubs is managed very efficiently that very little management overhead can be observed, and the users can focus more on different aspects of their projects.

2. Real time processing support:

Event Hubs has a special partition-based processing model which enables different applications process the same stream of the real time data at the same time. It enables capture of almost real-time data as micro-batches in processing which allows for longer retention of the captured data.

3. Scalability:

Event Hubs can manage and process data from a small data stream to a very huge data stream of Giga or Terabytes of data, so the problem of storage and processing overhead is out of question.

4. Ecosystem:

Event Hubs enables Kafka by Apache into the respective ecosystem to communicate with it seamlessly without having to manage clusters and immediately process the streams of data.

3.112 f4



title	f4
status	90
section	File systems
keywords	File systems

As the amount of data Facebook stores continues to increase, the need for quick access and efficient storage of data continues to rise. Facebook stores a class of data in Binary Large OBjects (BLOBs), which can be created once, read many times, never modified, and sometimes deleted. Haystack, Facebook's traditional BLOB storage system is becoming increasingly inefficient. The storage efficiency is measured in the effective-replication-factor of BLOBs.

f4 BLOB storage system provides an effective-replication-factor lower than that of Haystack. f4 is simple, modular, scalable, and fault tolerant. f4 currently stores over 65PBs of logical BLOBs, with a reduced effective-replication-factor from 3.6 to either 2.8 or 2.1 [269].

3.113 Facebook Corona



title	Facebook Corona
status	90
section	Cluster Resource Management
keywords	Cluster Resource Management

Corona is a new scheduling framework developed by facebook which separates the cluster resource management from job coordination. Facebook, employed the MapReduce implementation from Apache Hadoop since 2011 for job scheduling. The scheduling MapReduce framework has its limitations with the scalability as when the number of jobs at facebook grew in the next few years. Another limitation of Hadoop was it was a pull-based scheduling model as the task tracker have to provide a heartbeat to the job tracker to indicate that it is running which associated with a pre-defined delay, that was problematic for small jobs [270]. Hadoop MapReduce is also constrained by its static slot-based resource management model where a MapReduce cluster is divided into a fixed number of map and reduce slots based on a static configurations so the slots are not utilized completely anytime the cluster workload does not fit the static configuration.

Corona improves over the Hadoop MapReduce by introducing a cluster manager whose only purpose is to track the nodes in the cluster and the amount free resources [270]. A dedicated job tracker is created for each job and can run either in the same process as the client (for small jobs) or as a separate process in the cluster (for large jobs). The other difference is that it uses a push-based scheduling whose implementation does not involve a periodic heartbeat and thus scheduling latency is minimized. The cluster manager also implements a fair-share scheduling as it has access to the full snapshot of the cluster for making the scheduling decisions. Corona is used as an integral part of the Facebook's data infrastructure and is

helping power big data analytics for teams across the company.

3.114 Facebook Puma/Ptail/Scribe/ODS



title	Facebook Puma/Ptail/Scribe/ODS
status	90
section	Streams
keywords	Streams

The real time data Processing at Facebook is carried out using the technologies like Scribe, Ptail, Puma, and ODS. While designing the system, facebook primarily focused on the five key decisions that the system should incorporate which were Ease of Use, Performance, Fault-tolerance, Scalability, and Correctness.

"The real time data analytics ecosystem at facebook is designed to handle hundreds of Gigabytes of data per second via hundreds of data pipelines and this system handles over 200,000 events per second with a maximum latency of 30 seconds" [271].

Facebook focused on the Seconds of latency while designing the system and not milliseconds as seconds are fast enough to for all the use case that needs to be supported, and it allowed facebook to use persistent message bus for data transport and this also made the system more fault tolerant and scalable [271]. The large infrastructure of facebook comprises of hundreds of systems distributed across multiple data centers that needs a continuous monitoring to track their health and performance which is done by Operational Data Store (ODS) [272]. ODS comprises of a time series database (TSDB), which is a query service, and a detection and alerting system. ODS's TSDB is built atop the HBase storage system. Time series data from services running on Facebook hosts is collected by the ODS write service and written to HBase.

When the data is generated by the user from their devices, an AJAX

request is fired to facebook, and these requests are then written to a log file using Scribe (distributed data transport system), this messaging system collects, aggregates, and delivers high volume of log data with few seconds of latency and high throughput. Scribe stores the data in the HDFS (Hadoop Distributed File System) in a tailing fashion, where the new events are stored in log files and the files are tailed below the current events. The events are then written into the storage HBase on distributed machines. This makes the data available for both batch and real-time processing. Ptail is an internal tool built to aggregate data from multiple Scribe stores. It then tails the log files and pulls data out for processing. Puma is a stream processing system which is the real-time aggregation/storage of data. Puma provides filtering and processing of Scribe streams (with a few seconds delay), usually Puma batches the storage per 1.5 seconds on average and when the last flush completes, then only a new batch starts to avoid the contention issues, which makes it fairly real time.



title	Facebook Ta
status	10
section	NoSQL
keywords	NoSQL

Social giant Facebook is considered to be the largest warehouse of these social relationships. The consistency of massive data and large-scale read operations are technical difficulties. The traditional relational database model needs to be improved, and TAO comes into this revolution. On Facebook, people have formed a complex social network. In 2009, they began designing a new database architecture, TAO (The Associations and Objects). On June 25, 2013, Facebook officially announced to support its infrastructure details. [273].

Product engineers work between two completely different data models: large-scale MySQL servers use relational tables to store persistent data, and a similar number of cached data servers are used to store key-value pairs that SQL queries. Even the most common operations encapsulated in data access libraries, it still required product engineers to have a thorough understanding of the internals of the system to efficiently use the Memcache and MySQL. Tao is a data storage mechanism optimized for reading and deployed on Facebook as a single geographically distributed instance. Same as Google's Megastore, Spanner, it used MySql as back bone, and the upper Cache Server uses distributed Memcached.

TAO was originally designed to provide more than one billion read operations per second for a large data set. The TAO service runs on a large number of server clusters, and these geographically distributed clusters form a network. There are additional clusters for persistent storage of object and object associations, and RAM and flash for

caching. This hierarchical structure is more convenient when performing different types of cluster expansion alone, and can also effectively utilize server hardware. [274].

3.116 Facebook Tupperware



title	Facebook Tupperware
status	90
section	DevOps
keywords	DevOps

Facebook Tupperware is a system which provisions services by taking requirements from engineers and mapping them to actual hardware allocations using containers [275].Facebook Tupperware simplifies the task of configuring and running services in production and allows engineers to focus on actual application logic.The tupperware system consists of a Scheduler, Agent process and a Server Database. The Scheduler consists of set of machines with one of them as master and the others in standby.The machines share state among them.The Agent process runs on each and every machine and manages all the tasks and co-ordinates with the Scheduler.The Server database stores the details of resources available across machines which is used by the scheduler for scheduling jobs and tasks.Tupperware allows for sandboxing of the tasks which allows for isolation of the tasks.Initially isolation was implemented using chroots but now it is switched to Linux Containers (LXC) .The configuration for the container is done by a specific config file written in a dialect of python by the owner of the process.

3.117 Fiddler



title	Fiddler
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Fiddler is an HTTP debugging proxy server application. Fiddler captures HTTP and HTTPS traffic and logs it for the user to review by implementing man-in-the-middle interception using self-signed certificates. Fiddler can also be used to modify (fiddle with) HTTP traffic for troubleshooting purposes as it is being sent or received.[5] By default, traffic from Microsoft's WinINET HTTP (S) stack is automatically directed to the proxy at runtime, but any browser or Web application (and most mobile devices) can be configured to route its traffic through Fiddler [276].

3.118 FITS



title	FITS
status	10
section	File management
keywords	File management

FITS stand for ‘Flexible Image Transport System’. It is a standard data format used in astronomy. FITS data format is endorsed by NASA and International Astronomical Union. FITS can be used for transport, analysis and archival storage of scientific datasets and support multi-dimensional arrays, tables and headers sections [277]. FITS is actively used and developed [278]. FITS can be used for digitization of contents like books and magazines. Vatican Library uses FITS for long term preservation of their book, manuscripts and other collection [279]. Matlab, a language used for technical computing supports fits [280]. A 2011 paper explains how to perform processing of astronomical images on Hadoop using FITS [281].



title	Flink Streaming
status	10
section	Streams
keywords	Streams

Apache Flink is a framework for stateful computations over unbounded and bounded data streams [282]. Flink Streaming is a sub-component of the Apache Flink framework. Apache Flink allows users to analyze continuous data sources, however, Flink Streaming added usability by being able to process data streams in real time. Flink streaming framework allows it to process several different kinds of streaming data sources. Streaming data can be bounded, unbounded, real-time or recorded. Bounded streams have fixed data sets, whereas unbound do not. Apache Flink comes with the ability to process both types of streaming data. Along with this, streams can be real-time or recorded. Recorded streams are obviously easier to handle; however, Flink Streaming provides the ability to process data as it is generated. This is an extremely important part of the Apache Flink platform.

The DataStream API provided with Apache Flink provides primitives for many stream processing operations, these include: windowing, record-at-a-time transformation, and querying external data stores [282]. Flink Streaming comes with several features to allow streaming applications to remain functioning non-stop. Constant streaming of applications is bound to have failures occur, but Flinks robust feature packages allow these applications to continue functioning as if a failure has not occurred. Some of these features are: Consistent Checkpoints, Efficient Checkpoints, End-to-End Exactly Once, Integration with Cluster Managers, and High-Availability Setup.

Consistent Checkpoints and Efficient Checkpoints allow for

asynchronous checkpointing of the streaming application to eliminate the latency usually seen with stream checkpointing. This also provides fault tolerance for the continuously running applications. End-to-End Exactly Once, Integration with Cluster Managers, and High-Availability Setup allows for Apache Flink to maintain redundant backups to either restart failed processes immediately or eliminate single failure points. This occurs by duplication of processes so that if a failure occurs these duplicate processes can immediately start up. This allows for zero down time of the streaming applications.

3.120 Floe



title	Floe
status	10
section	Streams
keywords	Streams

3.121 Flume



title	Flume
status	90
section	Data Transport
keywords	Data Transport

Flume is distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of log data [283]. Flume was created to allow you to flow data from a source into your Hadoop environment. In Flume, the entities you work with are called sources, decorators, and sinks. A source can be any data source, and Flume has many predefined source adapters. A sink is the target of a specific operation. A decorator is an operation on the stream that can transform the stream in some manner, which could be to compress or uncompress data, modify data by adding or removing pieces of information, and more [284].

3.122 Foreman



title	Foreman
status	10
section	DevOps
keywords	DevOps



title	FTP
status	10
section	Data Transport
keywords	Data Transport

3.123.1 Old Text

FTP is an acronym for File Transfer Protocol [285]. It is network protocol standard used for transferring files between two computer systems or between a client and a server. It is part of the Application layer of the Internet Protocol Suite and works along with HTTP/SSH. It follows a client- server model architecture. Secure systems asks the client to authenticate themselves using a Username and Password registered with the server to access the files via FTP. The specification for FTP was first written by Abhay Bhushan in 1971 and is termed as RFC114 [286]. The current specification, RFC959 in use was written in 1985. Several other versions of the specification are available which provides firewall friendly FTP access, additional security extensions, support for IPV6 and passive mode file access respectively. FTP can be used in command line in most of the operating systems to transfer files. There are FTP clients such as WinSCP, FileZilla etc. which provides a graphical user interface to the clients to authenticate themselves (sign on) and access the files from the server.

3.123.2 New Text

The specification for FTP is defined in RFC114 Often FTP is used synonomosly as a protocoll specification as defined in [286] and as a service that uses the protocol to transfer files. For our discussion we refer to FTP as the service in the rest of the summary. FTP has been in use for nearly 3 decades. It is commonly used in both organizations and privately as an effective way to ensure that data files are

transported safely and accurately across some level of networking. FTP can be used in both Windows and Linux systems alike. Extension to FTP services are available that allow secure transmission of the data. Such extensins include SSL and it is know as sFTP ??? Transfer Protocols. To access the service, many operating systems offer the command

```
bash ftp source_file destination_directory
```

This command can be used with other flags that stipulate encryption, style of transportation, TCP port usage, and more.

"FTP is the user interface to the Internet standard File > Transfer Protocol. The program allows a user to transfer files > to and from a remote network site" ???.

For two connected devices, the simplicity of the command to transfer files is very effective for familiar *NIX users. FTP can also be used in a simplified console that includes many Unix basic commands, such as `mkdir` while permitting the same aforementioned functionality. For security purposes, usernames, and passwords are required and can be implemented in the command syntax, which is not recommended as password can appear in plain text in the operating system log files for command usages. Further functionality includes appending to files, omitting case sensitivity, modifying ownership, debugging, verifying hashes, setting timeout values, and more.

The File Transfer Protocol is essential to big data analytics because frequently used files need not be stored locally for usage using FTP. It is the backbone for many file transfer services and can alleviate the need of a Hadoop Distributed File System environment using effective scripting and permissions to enable access to required data. It also has important usage for data/disaster recovery in many corporations' storage environments. The data is ensured to be accurate once transferred and has appropriate security in place to ensure proper data usage across the systems in question.

3.124 FUSE



title	FUSE
status	10
section	File systems
keywords	File systems

FUSE (Filesystem in Userspace)

"is an interface for userspace programs to export a filesystem to the Linux kernel" [287].

The FUSE project consists of two components: the fuse kernel module and the libfuse userspace library. libfuse provides the reference implementation for communicating with the FUSE kernel module. The code for FUSE itself is in the kernel, but the filesystem is in userspace. As per a 2006 paper on HPTFS which has been built on top of FUSE [288]. It mounts a tape as normal file system based data storage and provides file system interfaces directly to the application. Another implementation of FUSE FS is CloudBB [289]. Unlike conventional filesystems CloudBB creates an on-demand two-level hierarchical storage system and caches popular files to accelerate I/O performance. On evaluating performance of real data-intensive HPC applications in Amazon EC2/S3, results show CloudBB improves performance by up to 28.7 times while reducing cost by up to 94.7% compared to the ones without CloudBB.

Some more implementation examples of FUSE are - mp3fs (A VFS to convert FLAC files to MP3 files instantly), Copy-FUSE (To access cloud storage on Copy.com), mtpfs (To mount MTP devices) etc.

3.125 Galaxy



title	Galaxy
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Ansible Galaxy is a website platform and command line tool that enables users to discover, create, and share community developed roles. Users' GitHub accounts are used for authentication, allowing users to import roles to share with the ansible community. A description of how Ansible roles are encapsulated and reusable tools for organizing automation content is available in [290]. Thus a role contains all tasks, variables, and handlers that are necessary to complete that role. Roles are depicted as the most powerful part of Ansible as they keep playbooks simple and readable [291].

"They provide reusable definitions that you can include whenever you need and customize with any variables that the role exposes."

GitHub hosts the project documents for Ansible Galaxy [292].

3.126 Galera Cluster



title	Galera Cluster
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Galera cluster is a type of database clustering which has all multiple masters and works on synchronous replication [293]. At a deeper level, it was created by extending MySQL replication API to provide all support for true multi master synchronous replication. This extended API is called as Write-Set Replication API and is the core of the clustering logic. Each transaction of wsrep API not only contains the record but also other meta-info to require to commit each node separately or asynchronously. So though it seems synchronous logically but works independently on each node. The approach is also called virtually synchronous replication. This helps in directly read-write on a specific node and can lose a node without handling any complex failover scenarios (zero downtime).

3.127 Galois



title	Galois
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Galois system was built by intelligent software systems team at University of Texas, Austin. Galois can be decibed as

"a system that automatically executes 'Galoized' serial C++ or Java code in parallel on shared-memory machinesv. It works by exploiting amorphous data-parallelism, which is present even in irregular codes that are organized around pointer-based data structures such as graphs and trees" [294].

By using Galois provided data structures programmers can write serial programs that gives the performance of parallel execution. Galois employs annotations at loop levels to understand correct context during concurrent execution and executes the code that could be run in parallel. The key idea behind Galois is Tao-analysis, in which parallelism is exploited at compile time rather than at run time by creating operators equivalent of the code by employing data driven local computation algorithm [295]. Galois currently supports C++ and Java.

3.128 Ganglia



title	Ganglia
status	90
section	Monitoring
keywords	Monitoring

Ganglia is a scalable distributed monitoring system for high-performance computing systems (clusters and grids). It is a BSD-licensed open-source project that grew out of the University of California, Berkeley Millennium Project which was initially funded in large part by the National Partnership for Advanced Computational Infrastructure (NPACI) and National Science Foundation RI Award EIA-9802069 [296].

It relies on a multicast-based listen/announce protocol to monitor state within clusters. It uses a tree of point-to-point connections amongst representative cluster nodes to unite clusters and aggregate their state [297]. It leverages technologies such as XML for data representation, XDR for compact, portable data transport, and RRDtool for data storage and visualization. The implementation is robust, has been ported to an extensive set of operating systems and processor architectures, and is currently in use on thousands of clusters around the world, handling clusters with 2000 nodes.

3.129 Genesis



title	Genesis
status	90
section	Interoperability
keywords	Interoperability



title	GFFS
status	10
section	File systems
keywords	File systems

GFFS [298] is an open, standards-based system, created for the need to access and manipulate remote resources such as file systems in a federated, secure, standardized, scalable and transparent manner, any third part data owners or application developers and users do not need to change their methods of storing or accessing data, The GFFS can employ global path-based namespace. Data in existing file systems, whether they are Windows file systems, MacOS file systems, AFS, Linux, or Lustre file systems can then be exported, or linked into the global namespace [298].

GFFS is easy to accommodate for any implementation. The first use of the GFFS at XSEDE is using the Genesis II implementation from the University of Virginia. Genesis II has been in continuous operation at the University of Virginia since 2007 in the Cross Campus Grid (XCG). In mid-2010, the XCG was extended to include FutureGrid resources at Indiana University, SDSC, and TACC [298].

GFFS is a fundamental component of the NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE) program. GFFS allows user applications to access (create, read, update, delete) remote resources in a location-transparent manner. Existing applications, whether static linked binaries, dynamically linked binaries or scripts (shell, PERL, Python), can access anywhere in GFFS without modification.

Transparent access to data is achieved by using operating system-specific file system drivers that understand the underlying standard

security, directory and file access protocols used by GFFS. These file system drivers map the GFFS global namespace to the local file system load, user can then access data and other resources in GFFS, just as you can access local files and directories [299].

3.131 Giraffe



title	Giraffe
status	90
section	Monitoring
keywords	Monitoring

Giraffe is a scalable distributed coordination service. Distributed coordination is a media access technique used in distributed systems to perform functions like providing group membership, gaining lock over resources, publishing, subscribing, granting ownership and synchronization together among multiple servers without issues. Giraffe was proposed as alternative to coordinating services like Zookeeper and Chubby which were efficient only in read-intensive scenario and small ensembles. To overcome this three important aspects were included in the design of Giraffe [300]. First feature is Giraffe uses interior-node joint trees to organize coordination servers for better scalability. Second, Giraffe uses Paxos protocol for better consistency and to provide more fault-tolerance. Finally, Giraffe also facilitates hierarchical data organization and in-memory storage for high throughput and low latency.



title	Giraph
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Apache Giraph is the graph processing tool from Apache org. Giraph is based on counterpart tool called Pregel, developed by Google. Apache Giraph is useful for processing large graphs of the real world, as it can analyze the trillion edges of the Facebook graph in 4 minutes [301].

Apache Giraph achieves these objectives by employing the Apache Hadoop cluster and a Bulk Synchronous Parallel Programming(BSP) technique. Apache Hadoop platform is primarily used for distributed processing of large datasets. It is designed to scale up to thousands of the machines forming cluster in no time [302].

On contrary, graph databases such as Neo4J are not performance efficient when dealing with large graphs. They mainly require random access to disk and are incapable of storing large graph data in memory. Furthermore, large graphs need to be partitioned, requires a distributed computation power and a parallel system. While Hadoop is famous for scalability, it was not mainly developed to process structural graph data. One of key disadvantages of Map-Reduce is that it is I/O intensive, and it needs to reload data for every iteration. Apache Giraph leverages Map-Reduce framework along with BSP model of distributed computation that runs parallel algorithms to boost performance for processing the large graphs [303].

Apache Giraph, essentially, is an iterative graph processing system. The typical Giraph job lifetime for processing graph information spans three phases, i.e., loading, compute, and offloading. The loading process loads the graph information into Giraph from the underlying file. This raw file contains the information about nodes and edges and may have various formats like JSON, text, etc. The compute phase is iterative, in which the worker nodes perform aggregation and update edge property based on the messages they receive and messages they have to forward. The process iterates until no messages are left to be processed. In the post-computation offloading phase, the results are written back to the file [304].

One of the famous examples of Apache Giraph for graph processing would be Google's web graph. Where each page represents a node, and the link connecting web pages to the page itself is considered to be an edge. Every page is then ranked using its popularity and importance in the graph. Another example would be the recommendation engines from social media platforms such as Facebook. In those platforms, each profile gets created from an individual's browsing history and interests. The engine then generates recommendations with other individuals who have similar interests and shows the same browsing trends, in a way the engine works as an iterative system to build a connection between like-minded people [304].

3.133 Gitreceive



title	Gitreceive
status	90
section	DevOps
keywords	DevOps

Gitreceive is used to create an ssh+git user which can accept repository pushes right away and also triggers a hook script. Gitreceive is used to push code anywhere as well as extend your Git workflow.

“Gitreceive dynamically creates bare repositories with a special pre-receive hook that triggers your own general gitreceive hook giving you easy access to the code that was pushed while still being able to send output back to the git user”

Gitreceive can also be used to provide feedback to the user not only just to trigger code on git push. Gitreceive can be used for the following:

“(a) for putting a git push deploy interface in front of App Engine (b) Run your company build/test system as a separate remote (c) Integrate custom systems into your workflow (d) Build your own Heroku e) Push code anywhere” [305].



title	Globus Online (GridFTP)
status	10
section	Data Transport
keywords	Data Transport

3.134.1 NEW TEXT

Globus online [306] is a data transfer management web service which enables to make a high-performance transfer of complex data. Data transfer occurs without interactions other systems and the process can restart itself if a problem occurs as well as automatically verify the integrity of the data after a transfer is complete. A Globus endpoint on the GridFTP server facilitates the process of data transfer using an online tool provided by Globus. An endpoint is a logical address on the GridFTP server that can be said to perform the role of a domain name in a web server. In Globus online, data is transferred between Globus endpoints. One can also communicate via a personal computer using Globus Connect Personal, which creates a personal endpoint to transfer data to and from the personal computer.

Globus transfer is easy and reliable. Many large institutions such as universities across the world rely on the Globus to handle their big data. Researchers can exchange big data securely on Globus. Since many individuals and multi-environment use Globus, it is the best platform to meet all types of people and exchange data. The Globus services are aligned to meet all the needs of the scientific research community across the world. Science is revolutionizing because of the ability of researchers to share data for both related and unrelated fields. I feel that Globus is enabling different branches of science to communicate in one language. For example, in cancer research, physics, biology, chemistry, mathematics, and other disciplines easily interchange data to steer research.

3.135 Globus Tools



title	Globus Tools
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

The Globus Toolkit is an open source toolkit organized as a collection of loosely coupled components [307]. These components consist of services, programming libraries and development tools designed for building Grid-based applications. GT components fall into five broad domain areas: Security, Data Management, Execution Management, Information Services, and Common Runtime [308]. These components enable a broader Globus ecosystem of tools and components that build on or interoperate with GT functionality to provide a wide range of useful application-level functions [309]. Since 2000, companies like Fujitsu, IBM, NEC and Oracle have pursued Grid strategies based on the Globus Toolkit [309].

3.136 Gluster



title	Gluster
status	10
section	File systems
keywords	File systems

3.137 Google and other public Clouds



title	Google and other public Clouds
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

A public cloud is a scenario where a provider provides services such as infrastructure or applications to the public over the internet. Google cloud generally refers to services such as cloud print, connect, messaging, storage and platform [310]. Google cloud print allows a print-aware application on a device, installed on a network, to provide prints to any printer on that network. Cloud connect allows an automatic storage and synchronization of Microsoft word documents, power-points and excel sheets to Google docs while preserving the Microsoft office formats. In certain cases, developers require important notifications to be sent to applications targeting android operating system. Google cloud messaging provides such services. Google cloud platform allows the developers to deploy their mobile, web and backend solutions on a highly scalable and reliable infrastructure [311]. It gives developers a privilege of using any programming language. Google cloud platform provides a wide range of products and services including networking, storage, machine learning, big data, authentication and security, resource management, etc. In general, public clouds provide services to different end users with the usage of the same shared infrastructure [312]. Windows Azure services platform, Amazon elastic compute cloud and Sun cloud are few examples of public clouds.

3.138 Google App Engine



title	Google App Engine
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Google App Engine is a cloud computing platform to host your mobile or web applications on Google managed servers. Google App Engine provides automatic scaling for web applications, i.e it automatically allocates more resources to the application upon increase in the number of requests. It gives developers the freedom to focus on developing their code and not worry about the infrastructure. Google App Engine provides built-in services and APIs such as load balancing, automated security scanning, application logging, NoSQL datastores, memcache, and a user authentication API, that are a core part to most applications [313].

An App Engine platform can be run in either the Standard or the Flexible environment. Standard environment lays restrictions on the maximum number of resources an application can use and charges a user based on the instance hours used. The flexible environment as the name suggests provides higher flexibility in terms of resources and is charged based on the CPU and disk utilization. The App Engine requires developers to use only its supported languages and frameworks. Supported languages are Java, Python, Ruby, Scala, PHP, GO, Node.js and other JVM oriented languages. The App Engine datastore uses a SQL like syntax called the GQL (Google Query Language) which works with non-relational databases when compared to SQL [314].



title	Google BigQuery
status	10
section	High level Programming
keywords	High level Programming

3.139.1 Old Text

Google BigQuery is an enterprise data warehouse used for large scale data analytics [315] [316]. A user can store and query massive datasets by storing the data in BigQuery and querying the database using fast SQL queries using the processing power of Google's infrastructure. In Google BigQuery a user can control access to both the project and the data based on his business needs which gives the ability to others to view and even query the data [316]. BigQuery can scale the database from GigaBytes to PetaBytes. BigQuery can be accessed using a Web UI or a command-line tool or even by making calls to the BigQuery REST API using a variety of client libraries such as Java, .NET or python. BigQuery can also be accessed using a variety of third party tools. BigQuery is fully managed to get started on its own, so there is no need to deploy any resources such as disks and virtual machines.

Projects in BigQuery are top-level containers in Google Cloud Platform [[www-bigquery-documentation](#)]. They contain the BigQuery Data. Each project is referenced by a name and unique ID. Tables contain the data in BigQuery. Each table has a schema that describes field names, types, and other information. Datasets enable to organise and control access to the tables. Every table must belong to a dataset. A BigQuery data can be shared with others by defining roles and setting permissions for organizations, projects, and datasets, but not on the tables within them. BigQuery stores data in the Capacitor columnar dataformat, and offers the standard

database concepts of tables, partitions, columns, and rows [317].

3.139.2 New Text

Google BigQuery [318] is intended as a data warehouse that streamlines access to large datasets without the need of the complicated intricacies that require DBAs. It is a RESTful information gathering tool that can use functionality that is similar to a SQL database. Because it is a Google product, it does require Google storage which allows it the benefit. Data is managed in a JSON form, allowing for powerful parser tools/libraries to be utilized in analysis [318].

This tool is effective for data analysts/experts who wish to use data that is highly scalable and require familiar functionality. Because it uses a SQL-like query language, on top of Google Storage, with all of its proprietary scripts and other analytics tools, it can be a cost-effective way for a small or large company to perform deep learning applications in simplified VDI and other company infrastructures [319].

One valuable feature is that separated compute and storage resources are available. This enables the use of cloud and other big data analytics by not requiring the storage of all data that is processed. This causes Google BigQuery to be the leader in a cause that all big data analysts require for their detailed analytics.

“BigQuery is the external implementation of one of the company’s core technologies whose code name is Dremel” ???.

Further features of the BigQuery tool include high data availability, real-time analytics/usage, local storage for foreign companies, governance for data integrity, and artificial/business intelligence and machine learning libraries [320]. It also includes other management tools such as Cost Controls and inventory/management consoles. There are many public datasets that can be used along with other proprietary options to further the cause of any customer using this

product. Its management is synchronous with processing which promotes comfort with the health status of the data in question [320]. As the storage is managed off-site, Google technicians are able to troubleshoot quickly with expertise that is otherwise unfound in big data storage and implementations.

3.140 Google Bigtable :Smiley: fa18-423-06



title	Google Bigtable
status	90
section	NoSQL
keywords	NoSQL

Google BigTable [321] is a NoSQL database service meant for large workloads and major analytical projects. BigTable is widely used by Google to hold petabytes of data for Google products and services such as Google Finance, Google Earth, and Google Analytics [322]. While performing like most databases, the designers of BigTable wanted to make it unique by adding focus on control of the data rather than layout and format [322]. Therefore, data collected is stored on one large table rather than multiple tables connected through relationships. The simple data model provides a focus on performance of collecting data simply.

"A Bigtable is a sparse, distributed, persistent multidimensional sorted map" [322].

It is setup to collect information that is stored from their services mentioned above. BigTable is designed to store information on many different versions of URLs, satellite imagery, and the many different services Google provides [323]. The structure of the data map that is indexed by a row key, column key, and a timestamp. The columns are grouped into column families that are used to group similar data types together [323]. The column families must be in place prior to storage of the data. A table is typically split into multiple collections of rows called tablets [322]. Tablets enable the table to have a sense of locality. Timestamps are used to create multiple versions of the same content.

To get an idea of the structure of the data in the table, look at this

example from an article by Paul Krzyzanowski [323]. CNN's website would be sorted into all the different pages and URLs that make up the website, these are the rows that would be stored in the data map. The column families would be the different languages that the webpages published in, the contents of the webpage, and an anchor. In reality, the data map would contain millions of columns within a column family. While the example is incredibly simple, it demonstrates the design of the data map that is used by Google for their services.



title	Google Chubby
status	10
section	Monitoring
keywords	Monitoring

Google Chubby is used as a lock service for a large number of small computing machines connected over a network. The main purpose of this lock service is to elect the primary among all the peers in a distributed system. Before Chubby, adhoc processes were used at google for the locking purpose which caused lot of manual intervention in case of any concerns. Chubby cells are created with all the servers connected in a distributed sytem and the applications will elect the server in a chuuby cell as the Master via RPC. Chubby cells are further sub-divided as replicas and for the application to elect a server as Master should obtain majority of votes from these replicas. The replicas will hold on to thier vote for a particular period of time. The usage of Google Chubby at Google helped the developers obtain the high availability of resources. Through-put and storage capacity were considered only after reliability and easily implementing capabilities in Google Chubby's design. It also helped to remove the handles on the servers that has mistakes with simple RPC commands. Chubby is

"Google's primary internal name service; it is a common rendezvous mechanism for systems such as MapReduce; the storage systems GFS and Bigtable use Chubby to elect a primary from redundant replicas; and it is a standard repository for files that require high availability, such as access control lists" [Mike Burrows, Google Inc].

Google Chubby combats asynchronous consensus which is solved by

the Paxos protocol. The implementation in Chubby is based on coarse grained lock server and a library that the client applications link against. As per the 2016 paper [324], an open-source implementation of the Google Chubby lock service was provided by the Apache ZooKeeper project. ZooKeeper used a Paxos-variant protocol Zab for solving the distributed consensus problem. Google stack and Facebook stack both use versions of zookeeper.

3.142 Google Cloud Dataflow



3.143 Google Cloud Dataflow

title	Google Cloud Dataflow
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Google Cloud Dataflow is a unified programming model and a managed service for developing and executing a wide variety of data processing patterns (pipelines). Dataflow includes SDKs for defining data processing workflows and a Cloud platform managed services to run those workflows on a Google cloud platform resources such as Compute Engine, BigQuery amongst others [325]. Dataflow pipelines can operate in both batch and streaming mode. The platform resources are provided on demand, allowing users to scale to meet their requirements, it's also optimized to help balance lagging work dynamically.

Being a cloud offering, Dataflow is designed to allow users to focus on devising proper analysis without worrying about the installation and maintaining the underlying data piping and process infrastructure [326].

3.143.1 Duplicated entry: merge

Google Cloud DataFlow is a unified programming model that manages the deployment, maintenance and optimization of data processes such as batch processing, ETL etc [327]. It creates a pipeline of tasks and dynamically allocates resources thereby maintaining high efficiency and low latency. These capabilities make it suitable for solving challenging big data problems [327]. Also, google DataFlow overcomes the performance issues faced by Hadoop Mapreduce while building pipelines [328]. The performance of MapReduce started deteriorating while facing multiple petabytes of

data whereas Google Cloud Dataflow is apparently better at handling enormous datasets [327]. Additionally Google Dataflow can be integrated with Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable, and BigQuery. The unified programming ability is another noteworthy feature which uses Apache Beam SDKs to support powerful operations like windowing and allows correctness control to be applied to batch and stream data processes.

3.144 Networking: Google Cloud DNS



title	Networking: Google Cloud DNS
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Under the umbrella of google cloud platform, helps user to publish their domain using Google's infrastructure. It is highly scalable, low latency, high availability DNS service residing on infrastructure same as google.

It is build around projects a resource container, domain for access control, and billing configuration. Managed zones holds records for same DNS name. The resource record sets collection holds current state of the DNS that make up managed zones it is unmodifiable or cannot be modified easily and changes to record sets. It supports A address records, AAAA IPv6, CAA Certificate authority, CNAME canonical name, MX mail exchange, NAPTR naming authority pointer, NS Name server record, SOA start of authority, SPF Sender policy framework, SRV service locator, TXT text record.

3.145 Google Cloud Machine Learning



title	Google Cloud Machine Learning
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Google Could Machi<ne Leaning is a Googles cloud based managed system for building machine learning model, capable to work on any type and volume of data. User can create their own machine learning model using GoogleTensorFlow framework, which helps to use the range of Google products from Google Photos to Google Cloud Speech. We can build our machine learning model regardless the size, google will managed it infrastructure according to requirement. User can immediately host the created model and start predicting on new data [329].Cloud Machine Learning provides two important things:

- Help user to train the machine learning model at large scale with the help of TensorFlow training application.
- User can host the trained model on cloud, this will help to use the large and new data available on cloud, which help in creating good model.
- Google CloudML will help user to focus on model instead of hardware configuration and resource management [330].



title	Google Cloud SQL
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Google Cloud SQL [331] or Cloud SQL is the database solution for developers looking to use a completely functional relational database in the cloud that is a part of the Google Cloud Platform (GCP). It is primarily being offered as a database that requires little management which is considered a big step forward as relational databases usually require a lot of attention from the administrators of the database. This helps the developers to focus on their business rather than spend more time managing the database. At the time of its launch Cloud SQL only came with the option of using MySQL [332] but now developers have the option of choosing between a MySQL instance or a PostgreSQL [333] instance and use Cloud SQL with it. There are a number of pricing options available for the developers to choose from for their application or business. These plans consist of various charges and vary depending on the relational database instance you choose between MySQL and PostgreSQL [334]. These charges can be paid per hour or per month. Regardless of which database instance you select for your business needs these are the common pricing charges that apply and there are some pricing charges which are database instance specific, the common pricing charges are as follows:

1. Instance Pricing: This charge depends on the machine you require, the amount of CPU power, RAM and storage capacity can be configured for these machines. The price is lower for a smaller machine and increases as you customize your machine. The charge is billed from the minute the machine is up and running.

2. Storage Pricing: This charge consists of the type of storage you choose for your machine; flash storage is more expensive when compared to the normal hard drive storage.
3. Network Pricing: The above-mentioned instance and storage charges also vary from region to region, if your machine is not in the same region, example: Inter-Continental then additional fee is applied.

3.147 Google Cloud Storage



title	Google Cloud Storage
status	10
section	File systems
keywords	File systems

Google Cloud Storage is the cloud enabled storage offered by Google [335]. It is unified object storage. To have high availability and performance among different regions in the geo-redundant storage offering. If you want high availability and redundancy with a single region one can go for Regional storage. Nearline and Coldline' are the different archival storage techniques. Nearline storage offering is for the archived data which the user access less than once a month. Coldline storage is the storage which is used for the data which is touched less than once a year.

All the data in Google Cloud storage belongs inside a project. A project will contains different buckets. Each bucket has different objects. We need to make sure that the name of the bucket is unique across all Google cloud name space. And the name of the objects should unique in a bucket.

3.148 Google DataStore



title	Google DataStore
status	90
section	NoSQL
keywords	NoSQL

Google Cloud Datastore is a NoSQL document database built for automatic scaling, high performance, and ease of application development [336]. Though Cloud Datastore interface has many of the features similar to traditional databases, but as a NoSQL database, it differs from the SQL in the way as it describes relationships between various data objects. It also provides a number of features that relational databases are not optimally suited to provide, including high-performance at a very large scale and high-reliability. The Google Cloud DataStore can have different kinds of properties for the same kind of entities, unlike the Relational Database where they are represented in rows. For example, the difference between entities can have the properties with the same name but having different values. The flexible schema maps naturally to object-oriented and scripting languages.

Non-relational databases have become popular recently, especially for web applications that require high-scalability and performance with high-availability. Non-relational databases such as Cloud DataStore let developers to choose an optimal balance between strong consistency and eventual consistency for each application. This allows developers to combine the benefits of both the database structures [337]. Datastore is designed to automatically scale to very large data sets, allowing applications to maintain high performance as they receive more traffic. Datastore also provides a number of features that relational databases are not optimally suited to provide, including high-performance at a very large scale and high-reliability [336].

3.149 Google Dremel



title	Google Dremel
status	90
section	High level Programming
keywords	High level Programming

Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, Google Dremel is capable of running aggregation queries over trillion-row tables in seconds [338]. With Dremel, you can write a declarative SQL-like query against data stored in a read-only columnar format efficiently for analysis or data exploration. It is also possible to write queries that analyze billions of rows, terabytes of data, and trillions of records in seconds. Dremel can be used for a variety of jobs including analyzing web-crawled documents, detecting e-mail spam, working through application crash reports.

3.150 Google F1



title	Google F1
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

F1 is a distributed relational database system built at Google to support the AdWords business. It is a hybrid database that combines high availability, the scalability of NoSQL systems like Bigtable, and the consistency and usability of traditional SQL databases. F1 is built on Spanner, which provides synchronous cross-datacenter replication and strong consistency [339].

F1 features include a strictly enforced schema, a powerful parallel SQL query engine, general transactions, change tracking and notification, and indexing, and is built on top of a highly-distributed storage system that scales on standard hardware in Google data centers. The store is dynamically sharded and is able to handle data center outages without data loss [340]. The synchronous cross-datacenter replication and strong consistency results in higher commit latency which can be overcome using hierarchical schema model with structured data types and through smart application design.

3.151 Google FlumeJava



o: flumeJava-parallel-pipelines is missing

title	Google FlumeJava
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

FlumeJava is a Java library that is built based on the concepts of MapReduce to simplify the development, testing, and execution of dataparallel pipelines [341].

FlumeJava is an easier-to-use version of MapReduce, making it simpler to build operation that process data. FlumeJava can also be integrated with other applications to allow stream processing instead of batch processing [342].

FlumeJava was able to optimize MapReduce tasks and decrease execution time of MapReduce by allowing roll-back failed job instead of restarting [343].

title	Google Fusion Tables
status	10
section	Application and Analytics
keywords	Application and Analytics

Google Fusion Tables, Google's cloud based webservice,[344] whose main purpose is to make management of data simpler, such that a novice can easily manage, share, visualize their data and collaborate with others while working on them was released in 2009 [345]. It also provides a platform to establishments which want to share their data privately, publicly or to users who want to collaborate across multiple enterprises [346].

Fusion tables has a REST API where the user can merge different tables from different sources. If needed a user can also make their data public thereby making it crawlable by search engines. Apart from integration, Google Fusion Tables provides a mirage of ways to visualize data in terms of different plots (e.g., scatter, bar etc.), charts, maps etc. and play with them. The visualizations provided by GFT work through Google Visualization API, once made a visualization can be embedded on any website by copying its javascript code fragment provided by GFT. The most famous of the visualizations provided by GFT is the Map. Users can add geographical information to their data and GFT uses google Maps to come with the respective visualizations [347].

Currently, Google lets 250 MB of data per dataset, and provides 1 GB quota per user . The data can be from variety of sources like Excel files (.xls, .xlsx, .ods), CSV files (Comma delimited), KML, TSV (any text delimited file) and also can chose data from public data available on Google Tables or millions of public tables from the web using GFT search engine. To encourage the data sharing, Google has many

mechanisms that offer incentives to the users. To improve collaboration experience, GFT acts as the platform where multiple users can discuss about the authenticity, meaning and correctness of data [346].

Storing and processing the massive amounts of data (tables, schemas, queries etc.) is a challenge faced by GFT, which has been tackled by having a sophisticated architecture in place to manage it. It is assembled on two layers of Google storage stack, which go by the name Bigtable and Megastore. Bigtable stores key value pairs, distributed among several servers based on key ranges, and every time a new table is made, it is added as a tuple automatically. It also stores the meta data of the tables such as transaction history in the tuples. Megastore is a library added above Bigtable which helps with maintaining consistent indexes, table transactions and replication of tables. The rows in a table are stored as one row in a single Bigtable that is dedicated to store all the user tables in GFT. Similarly, schemas of all user tables are stored in a Bigtable. In this way the Bigtable and Megastore help GFT have a scalable storage of data [347].

Being an experimental application from Google Research, GFT is the tool for fast and easy database management [346].



title	Google Kubernetes
status	90
section	DevOps
keywords	DevOps

Google Kubernetes is a cluster management platform developed by Google. Since 2014 Kubernetes has been open source and managed by The Cloud Native Computing Foundation [348]. Kubernetes is popular because of its flexibility and powerful capabilities to meet the demands of modern cloud-based architecture. Kubernetes is the result of efforts at Google to manage containers with hallmarks of both Infrastructure as a Service and Platform as Service. Building Kubernetes, google engineers had some specific goals;

"... make it easy to deploy and manage complex distributed systems, while still benefiting from the improved utilization that containers enable" [349].

So what is Kubernetes? In simple terms Kubernetes intended be the central platform and managing entity for applications, tools and workloads in any environment. Kubernetes is geared towards container environments where workloads go up and down. Kubernetes can organize and balance the connectivity, disk space and distributed computing in a containerized infrastructure [348].

Kubernetes is not limited to the cloud, but it seeks to be at the forefront of cloud architecture. Flexibility and extensibility are key Kubernetes hallmarks [350]. The goal with Kubernetes is to allow as many other solutions to be used in the infrastructure as possible. Kubernetes is intended to be extremely flexible. With this goal it doesn't dictate a CI or automation policy. It allows the user to build a container in a way that best serves the organizations needs then

provides the management tools to scale, manage, and maintain that complex cloud based applications or infrastructure [350].

Kubernetes can be broken into two architectural buckets; the master node and workers notes. Within the master node as you might expect one of the key components is the API server [351]. The API server controls the cluster and executes REST commands. Scheduled jobs and activities are initiated from the API server via the scheduler. Outside of the master node are the worker nodes Pods are run in the worker nodes [351].Pods are run on the same host and can contain a group of containers that work together. Pods share volumes and network connectivity and other resources. Kubelets receive from the API server configuration info for the pod.

3.154 Google MillWheel



title	Google MillWheel
status	90
section	Streams
keywords	Streams

MillWheel is a framework for building low-latency data-processing applications. Users specify a directed computation graph and application code for individual nodes, and the system manages persistent state and the continuous flow of records, all within the envelope of the framework's fault-tolerance guarantees. Other streaming systems do not provide this combination of fault tolerance, versatility, and scalability. MillWHEEL allows for complex streaming systems to be created without distributed systems expertise. MillWheel's programming model provides a notion of logical time, making it simple to write time-based aggregations. MillWheel was designed from the outset with fault tolerance and scalability in mind. In practice, we find that MillWheel's unique combination of scalability, fault tolerance, and a versatile programming model [352].

3.155 Google Omega



title	Google Omega
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

3.156 Google Prediction API & Translation API



title	Google Prediction API & Translation API
status	90
section	Application and Analytics
keywords	Application and Analytics

Google Prediction API & Translation API are part of Cloud ML API family with specific roles. Below is a description of each and their use.

Google Prediction API provides pattern-matching and machine learning capabilities. Built on HTTP and JSON, the prediction API uses training data to learn and consecutively use what has been learned to predict a numeric value or choose a category that describes new pieces of data. This makes it easier for any standard HTTP client to send requests to it and parse the responses. The API can be used to predict what users might like, categorize emails as spam or non-spam, assess whether posted comments sentiments are positive or negative or how much a user may spend in a day. Prediction API has a 6 month limited free trial or a paid use for \$10 per project which offers up to 10,000 predictions a day [353].

Google Translation API is a simple programmatic interface for translating an arbitrary string into any supported language. Google Translation API is highly responsive allowing websites and applications to integrate for fast dynamic translation of source text from source language to a target language. Translation API also automatically identifies and translate languages with a high accuracy from over a hundred different languages. Google Translation API is charged at \$20 per million characters making it an affordable localization solution. Translation API is also distributed in two editions, premium edition which is tailored for users with precise long-form translation services like livestream, high volumes of emails or detailed articles and documents. There's also standard edition

which is tailored for short, real-time conversations [354].

3.157 Google Pub Sub



title	Google Pub Sub
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Google Pub/Sub provides an asynchronous messaging facility which assists the communication between independent applications [355]. It works in real time and helps keep the two interacting systems independent. It is the same technology used by many of the Google apps like GMail, Ads, etc. and so integration with them becomes very easy. Some of the typical features it provides are: (1) Push and Pull - Google Pub/Sub integrates quickly and easily with the systems hosted on the Google Cloud Platform thereby supporting one-to-many, one-to-one and many-to-many communication, using the push and pull requests. (2) Scalability - It provides high scalability and availability even under heavy load without any degradation of latency. This is done by using a global and highly scalable design. (3) Encryption - It provides security by encryption of the stored data as well as that in transit. Other than these important features, it provides some others as well, like the usage of RESTful APIs, end-to-end acknowledgement, replicated storage, etc [356].

3.158 Gora (general object from NoSQL)



title	Gora (general object from NoSQL)
status	10
section	In-memory databases/caches
keywords	In-memory databases/caches

Gora is a in-memory data model which also provides persistence to the big data [357]. Gora provides persistence to different types of data stores. Primary goals of Gora are:

1. data persistence
2. indexing
3. data access
4. analysis
5. map reduce support

Unlike ORM models which mostly work with relational databases for example hibernate gora works for most type of data stores like documents, columnar, key value as well as relational. Gora uses beans to maintain the data in-memory and persist it on disk. Beans are defined using apache avro schema. Gora provides modules for each type of data store it supports. The mapping between bean definition and datastore is done in a mapping file which is specific to a data store. Type Gora workflow will be:

1. define the bean used as model for persistence
2. use gora compiler to compile the bean
3. create a mapping file to map bean definition to datastore
4. update gora.properties to specify the datastore to use
5. get an instance of corresponding data store using datastore factory.

Gora has a query interface to query the underlying data store. Its configuration is stored in gora.properties which should be present in classpath. In the file you can specify default data store used by Gora engine. Gora also has a CI/CD library call GoraCI which is used to write integration tests.

3.159 Granules



title	Granules
status	90
section	Streams
keywords	Streams

Granules are used for execution or processing of data streams in distributed environment. When applications are running concurrently on multiple computational resources, granules manage their parallel execution. The MapReduce implementation in Granules is responsible for providing better performance. It has the capability of expressing computations like graphs. Computations can be scheduled based on periodicity or other activity. Computations can be developed in C, C++, Java, Python, C#, R. It also provides support for extending basic Map reduce framework. Its application domains include hand writing recognition, bio informatics and computer brain interface [358].

3.160 GraphBuilder (Intel)



title	GraphBuilder (Intel)
status	90
section	Application and Analytics
keywords	Application and Analytics

Intel GraphBuilder for Apache Hadoop V2 is a software that is used to build graph data models easily enabling data scientists to concentrate more on the business solution rather than preparing/formatting the data. The software automates (a) Data cleaning, (b) transforming data and (c) creating graph models with high throughput parallel processing using hadoop, with the help of prebuilt libraries. Intel Graph Builder helps to speed up the time to insight for data scientists by automating heavy custom workflows and also by removing the complexities of cluster computing for constructing graphs from Big Data. Intel Graph Building uses Apache Pig scripting language to simplify data preparation pipeline.

"Intel Graph Builder also includes a connector that parallelizes the loading of the graph output into the Aurelius Titan open source graph database - which further speeds the graph processing pipeline through the final stage".

Finally being an open source there is a possibility of adding a load of functionalities by various contributors [359].

3.161 GraphChi



title	GraphChi
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

GraphChi is a disk-based system for computing efficiently on graphs with large number of edges. It uses a well-known method to break large graphs into small parts, and executes data mining, graph mining, machine learning algorithms. GraphChi can process over one hundred thousand graph updates per second, while simultaneously performing computation [360]. GraphChi is a spin-off of the GraphLab. GraphChi brings web-scale graph computation, such as analysis of social networks, available to anyone with a modern laptop

3.162 graphdb



title	graphdb
status	10
section	NoSQL
keywords	NoSQL

A Graph Database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data [361]. The Graph is a concept which directly relates the data items in the store. The data which is present in the store is linked together directly with the help of relationships. It can be retrieved with a single operation. Graph database allow simple and rapid retrieval of complex hierarchical structures that are difficult to model in relational systems.

There are different underlying storage mechanisms used by graph databases. Some graphdb depend on a relational engine and store the graph data in a table, while others use a key-value store or document-oriented database for storage. Thus, they are inherently caled as NoSQL structures. Data retrieval in a graph database requires a different query language other than SQL. Some of the query languages used to retrieve data from a graph database are Gremlin, SPARQL, and Cypher. Graph databases are based on graph theory. They employ the concepts of nodes, edges and properties.



title	GraphLab
status	10
section	Application and Analytics
keywords	Application and Analytics

GraphLab [362] is a graph-based parallel framework for C++, especially for machine learning.

"GraphLab was initially an academic project" [363].

Carlos Guestrin is the co-founder of Dato, which was previously known as GraphLab. Students of Guestrin were working on large scale algorithms. After they tried to implement those large scale algorithms on top of Hadoop, they found the running speed would be really slow. Then, they decided to make a system that could save them a lot of times on this, which is the origin of GraphLab [364].

Given the scale and complexity of real world data, the GraphLab has been designed by incorporating different high-level algorithms for a higher performance [365]. With algorithms like Stochastic Gradient Descent incorporated, GraphLab can help create or install large scale applications in a convenient way. Furthermore, GraphLab comprises toolkits for machine learning, and neat libraries for data transformation, manipulation and model visualization, which is the key reason why GraphLab is helpful and easy to use. There are three main CREATE architectures for GraphLab: SFrame, SGraph and Machine Learning [365]. One of the benefits that using GraphLab brings, which is also the original reason why GraphLab was built, is that it can handle large data, with help of SFrame and SGraph.

SFrame, one of the data structures used in GraphLab, is a disk-based tabular data structure that

“helps to scale analysis and data processing to handle large data set (Tera byte), even on your laptop” [365].

Also, the SFrame syntax is similar to the widely-used library pandas. SFrame collect elements stored on disk and make them into SArray columns [365]. Then, SGraph object is used in GraphLab to help perform a graph-oriented data analysis, which presents items as vertex and relationships between items as edges. After that, GraphLab has a browser-based interactive GUI call GraphLab Canvas. It makes users be able to

“explore tabular data, summary statistics and bi-variate plots” [365],

which help users save time coding.

3.164 GraphX



title	GraphX
status	90
section	Application and Analytics
keywords	Application and Analytics

GraphX is Apache Spark's API for graph and graph-parallel computation [366].

GraphX provides:

Flexibility: It seamlessly works with both graphs and collections. GraphX unifies ETL, exploratory analysis, and iterative graph computation within a single system. You can view the same data as both graphs and collections, transform and join graphs with RDDs efficiently, and write custom iterative graph algorithms using the Pregel API.

Speed: Its performance is comparable to the fastest specialized graph processing systems while retaining Apache Spark's flexibility, fault tolerance, and ease of use.

Algorithms: GraphX comes with a variety of algorithms such as PageRank, Connected Components, Label propagations, SVD++, Strongly connected components and Triangle Count.

It combines the advantages of both data-parallel and graph-parallel systems by efficiently expressing graph computation within the Spark data-parallel framework [[@www-graphX1](#)].

It gets developed as a part of Apache Spark project. It thus gets tested and updated with each Spark release.

3.165 Graylog



title	Graylog
status	90
section	Application and Analytics
keywords	Application and Analytics

Graylog is an open source log management tool that allows an organization to assemble, organize and analyze large amounts of data from its network activity. It collects and aggregates events from a group of sources and presents data in a streamlined, simplified interface where one can drill down to significant metrics, identify key relationships, generate powerful data visualizations and derive actionable insights [367]. Graylog allows us to centrally collect and manage log messages of an organization's complete infrastructure [368]. A user can perform search on terabytes of log data to discover number of failed logins, find application errors across all servers or monitor the activity of a suspicious user id. Graylog works on top of ElasticSearch and MongoDB to facilitate this high availability searching. Graylog provides visualization through creation of dashboards that allows a user to build pre-defined views on his data to assemble all of his important data only a single click away [369]. Any search result or metric shall be added as a widget on the dashboard to observe trends in one single location. These dashboards can also be shared with other users in the organization. Based on a user's recent search queries, graylog also allows you to distinguish data that are not searched upon very often and thus can be archived on cost effective storage drives. Users can also add certain trigger conditions that shall alert the system about performance issues, failed logins or exceptions in the flow of the application.



o: quotation preferred

title	H-Store
status	10
section	In-memory databases/caches
keywords	In-memory databases/caches

H-Store was an experimental database management system which just got its final release in June, 2016. It was initially developed by database researchers from CMU, MIT, Yale and Brown University and was funded by Intel in 2007. It is written in C++ and Java and is available on Linux and MacOS operating systems [370]. H-store was particularly designed for Online Transactional Processing to mitigate various problems faced by traditional relational database systems (RDBMS) while processing repetitive short transactions which are a part of the workload processed by an OLTP system. H-store introduces the distributed relational database which is row based. It promotes a parallel DBMS which provides high performance of the NoSQL database but still holds the reliability of a traditional DBMS system [371]. H-store runs on a cluster deployed on the same domain but has two or more than two nodes or physical computers to perform the transactional computations. Parameterized SQL commands along with control code is stored in stored procedures. These stored procedures are called each time an OLTP application needs to make a call to the database system powered by H-Store. During deployment a framework with stored procedures, database schema and workload is provided to the administrator. These stored procedures can be referenced using unique invocations at runtime. It is advisable to introduce all stored procedures at deployment phase only as if new procedures are added, the database design won't be optimized for the new procedures. Whenever an OLTP transaction is used, the application then needs to invoke the uniquely referenced stored procedures provided at the time of deployment using

parameters that can be passed by the client as input. These applications can be executed by any node in the H-store cluster irrespective of the fact that the data is stored in that node or some other node. Using this deployment H-store is able of optimize some of the features of OLTP applications such as short set of transactions and repeated calling of the stored procedures [372].

3.167 H2O



title	H2O
status	90
section	Application and Analytics
keywords	Application and Analytics

It is an open source software for big data analysis. It was launched by the Start-up H2O in 2011. It provides an in-memory, distributed, fast and a scalable machine learning and predictive analytics platform that allows the users to build machine learning models on big data [373]. It is written in Java. It is currently implemented in 5000 companies. It provides APIs for R (3.0.0 or later), Python (2.7.x, 3.5.x), Scala (1.4-1.6) and JSON [374]. The software also allows online scoring and modeling on a single platform. It is scalable and has a wide range of OS and language support. It works perfectly on the conventional operating systems, and big data systems such as Hadoop, Cloudera, MapReduce, HortonWorks. It can be used on cloud computing environments such as Amazon and Microsoft Azure [375].

3.168 Hadoop



title	Hadoop
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Apache Hadoop is an open source framework written in Java that utilizes distributed storage and the MapReduce programming model for processing of big data. Hadoop utilizes commodity hardware to build fault tolerant clusters. Hadoop was developed based on papers published by Google on the Google File System (2003) and MapReduce (2004) [376].

Hadoop consists of several modules: the Cluster, Storage, Hadoop Distributed File System (HDFS) Federation, Yarn Infrastructure, MapReduce Framework, and the Hadoop Common Package. The Cluster is comprised of multiple machines, otherwise referred to as nodes. Storage is typically in the HDFS. HDFS federation is the framework responsible for this storage layer. YARN Infrastructure provides computational resources such as CPU and memory. The MapReduce layer is responsible for implementing MapReduce [377]. The Hadoop Common Package which includes operating and file system abstractions and JAR files needed to start Hadoop [376].

3.169 HadoopDB



title	HadoopDB
status	90
section	High level Programming
keywords	High level Programming

HadoopDB is a hybrid of parallel database and MapReduce technologies. It approaches parallel databases in performance and efficiency, yet still yields the scalability, fault tolerance, and flexibility of MapReduce systems. It is a free and open source parallel DBMS. The basic idea behind it is to give Hadoop access to multiple single-node DBMS servers (eg. PostgreSQL or MySQL) deployed across the cluster. It pushes as much as possible data processing into the database engine by issuing SQL queries which results in resembling a shared-nothing cluster of machines [378].

HadoopDB is more scalable than currently available parallel database systems and DBMS/MapReduce hybrid systems. It has been demonstrated on clusters with 100 nodes and should scale as long as Hadoop scales, while achieving superior performance on structured data analysis workloads.

3.170 Hama



title	Hama
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Apache Hama is a framework for Big Data analytics which uses the Bulk Synchronous Parallel (BSP) computing model, which was established in 2012 as a Top-Level Project of The Apache Software Foundation. It provides not only pure BSP programming model but also vertex and neuron centric programming models, inspired by Google's Pregel and DistBelief [379]. It avoids the processing overhead of MapReduce approach such as sorting, shuffling, reducing the vertices etc. Hama provides a message passing interface and each superstep in BSP is faster than a full job execution in MApReduce framework, such as Hadoop [380].

3.171 Harp



title	Harp
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Harp is a simple, easy to maintain, low risk and easy to scale static web server that also serves Jade, Markdown, EJS, Less, Stylus, Sass, and CoffeeScript as HTML, CSS, and JavaScript without any configuration and requires low cognitive overhead [381]. It supports the beloved layout/partial paradigm and it has flexible metadata and global objects for traversing the file system and injecting custom data into templates. It acts like a lightweight web server that was powerful enough for me to abandon web frameworks for dead simple front-end publishing. Harp can also compile your project down to static assets for hosting behind any valid HTTP server.

3.172 Haystack



title	Haystack
status	10
section	File systems
keywords	File systems

Haystack is an open source project working with data from internet of Things, aim to standardise the semantic data model generated from smart devices, homes, factories etc. It include automation, control, energy, HVAC, lighting and other environmental systems [382].

Building block of Project haystack is on TagModel tagging of metadata stored in key/value pair applied to entity such id, dis, sites, geoAddr, tz. Structure the primary structure of haystack is based on three entities, Site location of single unit, equip physical or logical piece of equipment within site, point sensor, actuator or setpoint value for equip, it also includes weather outside weather condition. TimeZone time series data is most important factor it is foundation for sensor and operational data. Captured data not always associated with measurable unit, however it provides facility to associate the data points. Commonly Supported units like Misc, Area, Currency, Energy, Power, Temperature, Temperature differential, Time, Volumetric Flow. The data often represented in 2D tabular form for tagged entities. It supports the query language for filtering over the data, data exposed through REST API in JSON format.

3.173 Hazelcast



title	Hazelcast
status	90
section	In-memory databases/caches
keywords	In-memory databases/caches

Hazelcast is a java based, in memory data grid [383]. It is open source software, released under the Apache 2.0 License [384]. Hazelcast enables predictable scaling for applications by providing in memory access to data. Hazelcast uses a grid to distribute data evenly across a cluster. Clusters allow processing and storage to scale horizontally. Hazelcast can run locally, in the cloud, in virtual machines, or in Docker containers. Hazelcast can be utilized for a wide variety of applications. It has APIs for many programming languages including Python, Java, Scala, C++, .NET and Node.js and supports any binary languages through an Open Binary Client Protocol [383].



title	HBase
status	10
section	NoSQL
keywords	NoSQL

HBase [385] is a data storing system especially for large and extremely sizable data which is stored in Hbase clusters. Contrary to traditional Relational Database Management Systems(RDBMS), HBase has features that is similar to Non-Relational Database where it is unable to use Structure Query Language (SQL) and hence it is known for its data storing as opposed to database. As the feature differs, HBase is exceptionally efficient in terms of reading and writing capabilities due to its auto sharding system which allow data stored in multiple Region Servers in such a way that it can scale horizontally in conjunction with the regions ???.

The use of HBase is more beneficial for large quantity data, i.e. peta byte - billions of rows - as compare to small volume in which case using other database would be more efficient. This is due to the nature of the nodes scalability where its individual node is able to perform relatively well for medium size data, thousands and/or millions of rows. As such, running on small scale will result in inactivity of other nodes which defeat the purpose of utilizing region servers. On similar note, the machines required will also have the same proportionality in correspond to the number of nodes for the system to work efficiently ??? [386].

The data stored in HBase can be structured differently depending on the usage. However, all table needs to have primary keys which compartmented into column families. These column families are assigned into different RegionServers which controlled by HMaster. As a column-oriented database, Hbase usage prominently on real-

time and random data which is more suitable for “Online Analytical Processing” than its counter parts, “Online Transactional Processing” [387]. HBase also allows vertical and horizontal split of tables which served different purposes. While vertical split stored information into separate files in the same regions, horizontal split is the default configuration where columns of tables are stored in multiple regions [388] [389].



title	HCatalog
status	10
section	High level Programming
keywords	High level Programming

HCatalog [390] is a metadata repository that allows you to share metadata across different sources no matter where the data resides or in what format it is stored. It is a table storage management tool for Hadoop. HCatalog makes tabular data of Hive metastore available to other Hadoop applications. It facilitates users to write the data onto the grid, no matter which data processing tool they use. It provides a consistent interface between Apache Hive, Apache Pig, and MapReduce [390]. It can be considered as an extension of the hive because it ships with Hive. File paths are not required for HCatalog. HCatalog supports file formats like RCFile, CSV, JSON, SequenceFile, ORC

The main aim of HCatalog is to allow Pig and MapReduce to use the same data structures as those used by hive so that there is no need for data conversion. Hive stores its metadata in Derby or MySQL. The other two do that using a code written into programs and output and input operations. We can write a serializer-deserializer in HCatalog that provisions writing and reading of files in any format. It makes sure that users do not have to stress about where the data is stored or in what format the data is stored.

Hadoop opens up a lot of opportunity for the enterprise as a processing and storage environment. HCatalog integrates Hive's DDL and is developed with Hive metastore as base. HCatalog loader accepts a table to write to and also optionally a selection predicate to indicate which of the partitions to scan. HCatalog uses its Command Line Interface to define data, therefore, it doesn't rely on MapReduce

for querying. The data presented by HCatalog is relational in view. The tables are either hash partitioned on one or more keys that means for a given value of the key there will be one partition that contains all rows [391].



title	HDF
status	10
section	File management
keywords	File management

HDF (Hortonworks DataFlow) provides users a GUI based platform to design and build complex dataflows to ingest and analyze data from multiple sources of streaming data [392]. In an age where Internet of things is rapidly gaining importance with each passing day, systems ability to deal with a large amount of streaming data has become paramount. But dealing with streaming data brings with it, a lot of challenges. Some of the main aspects while dealing with real-time data include - Data security, Computational Speed and maintaining Data Integrity. There are a few technologies developed specifically to handle streaming data. Hortonworks DataFlow (HDF) is one such technology.

“Hortonworks DataFlow (HDF) is a scalable, real-time streaming analytics platform that ingests, curates and analyzes data for key insights and immediate actionable intelligence” [392].

While ingesting data in HDF, it is possible for the users to transform and enrich the data as well. HDF uses Apache Kafka to perform real-time analytics on extremely large amounts of the streaming data, enabling the users to make faster decision. HDF is an open source technology, making the setup future-proof. Its ability to handle a large volume of diversified data means that HDF is used in multiple sectors to implement IOT solutions. Real-time dataflow management and provenance eliminates the need to perform any manual sweep to look for any missing data or any shortcomings in the dataflow. HDF

provides what can be called an end-to-end solution while dealing with streaming data; which means that, it provides a solution beginning right from data collection, data security, data transformation, integration from multiple sources and then tops it all off by deploying Machine learning algorithms to provide actionable business insights to the end user and it does all of this in real-time. With GDPR and other regulatory compliance laws, tracking of data lineage has become a compulsion; HDFs integration of Apache Atlas provides the user with complete control over data governance [392].

3.177 HDFS



title	HDFS
status	90
section	File systems
keywords	File systems

Hadoop provides distributed file system framework that uses Map reduce (Distributed computation framework) for transformation and analyses of large dataset. Its main work is to partition the data and other computational tasks to be performed on that data across several clusters. HDFS is the component for distributed file system in Hadoop. An HDFS cluster primarily consists of a Name Node and Data Nodes. Name Node manages the file system metadata such as access permission, modification time, location of data and Data Nodes store the actual data. When user applications or Hadoop frameworks request access to a file in HDFS, Name Node service responds with the Data Node locations for the respective individual data blocks that constitute the whole of the requested file [393].



title	Helix
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

Helix is a cluster management tool developed by Apache. We run a distributed system on multiple nodes as it provides us scalability, fault tolerance and ability of load balancing in case of a fault so that the system performance doesn't suffer. In a cluster managed distributed system, the primary functions performed by the nodes include data storage and producing data streams to be used further etc. Helix makes decisions for a cluster-based system which requires end to end coordination between the clusters and knowledge of the entire architecture. Helix enables an automatic management of clusters by replicating and partitioning the resources that are required by each cluster. Helix models a distributed system using the concepts of Automated Finite State machines and transitions [394]. It enables fault reporting and takes part in the recovery. Helix monitors the overall health of the cluster and gives out timely alerts and provides automatic load balancing which is extremely important in today's IT because of the strict adherence to the SLAs. It reacts to any changes that occur within the system and comes up with a plan to bring the system back to initial stable condition. Helix also has a centralized configuration management, so an administrator doesn't need to modify the configuration at every node at the time of deployment. Helix is being widely used in the LinkedIn ecosystem to manage its backend system Espresso, instant messaging feature of Instagram, Uber uses Helix extensively in the data delivery system to move data to Kafka, HBase and HDFS. Turn uses Helix in the key value store to manage partitions and automate migration process in case there is a change in the number of machines in a cluster [395]. Currently more work is being done to allow Helix to perform task

management operations and integrate with other resource management systems.



title	Heroku
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Heroku's main goal is to support customer focused apps and is a cloud-based service. Heroku provides many simple, easy to use and efficient tools hence becoming a top PaaS provider. Heroku now supports Ruby, Java, Scala, Node.js, Python and Clojure [396].

The projects that use Heroku does not need infrastructure support as the platform also manages the hardware and servers. Several programming languages are supported by Heroku. Polymorphism and scalability are what makes Heroku preferable for smaller projects, Heroku uses a service model which is free and uses tiered service packages for complex projects. It is based on a managed container called dynos. The integration of the data service is done with Heroku's ecosystem. This data service enables the running of the modern applications [397].

Heroku has several add-ons. To analyze hosted applications logs, its events etc, Heroku uses one of its add-on which is called the Treasure Data Hadoop.

"Treasure Data toolbelt can be installed along with Heroku plugin to manage treasure Data Hadoop on Heroku" [398].

Treasure data collects, stores and analyzes large amount of data immediately. This also helps in maintaining a log of the events. Typical use cases are for:

- Conversion path analysis

- Ranking calculation
- Reports of the customer etc [399].

With Heroku any application can be deployed to the cloud with one push from Git. Today, Heroku has expanded to offer Heroku Enterprise platform and a Heroku Postgres SQL database as a service for managing large amount of data on cloud. Deployment of the applications in Heroku happens in the below mentioned ways [400]:

- Prepping: Heroku takes the applications source code, dependency description (instructions that is required for the application to run), and a file that provides the process method of the app (procfile) which also exposes the architectural components of the applications. These are the only components needed to build an application and to generate an executable file in Heroku.
- Exporting code Git: Heroku starts the build by pushing the code to Git using a simple command.
- Build Phase: During this phase the code is compiled, and the output is generated.
- Execution Phase: Heroku uses dynos which is a mini operating system, which handles the applications.
- The release: The release is the final product that Heroku delivers [400].

3.180 Hibernate



title	Hibernate
status	90
section	Object-relational mapping
keywords	Object-relational mapping

Hibernate is an open source project which provides object relational persistence framework for applications in Java. It is an Object relational mapping library (ORM) which provides the framework for mapping object oriented model to relational database. It provides a query language, a caching layer and Java Management Extensions (JMX) support. Databases supported by Hibernate includes DB2, Oracle, MySQL, PostgreSQL. To provide persistence services, Hibernate uses database and configuration data. For using hibernate, firstly a java class is created which represents table in the database. Then columns in database are mapped to the instance variables of created Java class. Hibernate can perform database operations like select, insert, delete and update records in table by automatically creating query. Connection management and transaction management are provided by hibernate. Hibernate saves development and debugging time in comparison to JDBC. But it is slower at runtime as it generates many SQL statements at runtime. It is database independent. For batch processing it is advisable to use JDBC over Hibernate [401].

3.181 Hive fa18-423-05



title	Hive
status	10
section	High level Programming
keywords	High level Programming

Apache Hive [402] is a data warehouse software project built on the top of Apache Hadoop.

“It is an Open Source data warehousing system. It is exclusively used to query and analyze huge data sets stored in the Hadoop storage” [403].

More exhaustively, Hive is part of the software ecosystem based on Hadoop framework, mainly for

“handling large data sets in a distributed computing environment” [404].

The intention of Hive is to simplify the process of operating the MapReduce framework, which requires users have advanced understanding of Java programming, and to open Hadoop for a wider group of users [404].

There are three important functionalities of Hive. They are separately data analysis, data query and data summarization [403]. For querying, the only language supported by Hive is the HiveQL. Basically, HiveQL can translate queries into MapReduce jobs for deploying, with supports to MapReduce scripts. In other words, what Hive does for returning the query value is to convert particular SQL query into MapReduce job, before submitting which to the Hadoop cluster[405]. As a result,

“the partitioning process decreases the operational I/O

time and decreases execution load" [405],

which leads to the increasing in the overall performance. There is another function called Dynamic Runtime Filtering helps saving CPU occupancy and network consumption: a filter would work on actual dimension table values and it will eliminate rows that do not match requirements, which would be skipped for further operations like shuffling [405].

Above all, Hive has the advantage that while managing large data sets, it still keep a high data reading and writing speed [403]. Typically, comparing to other queries, with the same type of large data sets, Hive would have a faster response time. Besides advantages mentioned above, Hive is also flexible since there is an increasing number of commodities can be easily added

"in response to more adding of cluster of data without any drop in performance" [403].

3.182 (a) publish-subscribe: MPI



title	(a) publish-subscribe: MPI
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

see \URL{<http://www.slideshare.net/Foxsden/high-performance-processing-of-streaming-data>}

3.183 HPX-5



title	HPX-5
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

High Performance ParallelX (Hpx-5) is an open source, distributed model that provides opportunity for operations to run unmodified on one-to-many nodes [406]. The dynamic nature of the model accommodates effective

“computing resource management and task scheduling”.

It is portable and performance-oriented. HPX-5 was developed by IU Center for Research in Extreme Scale Technologies (CREST). Concurrency is provided by lightweight control object (LCO) synchronization and asynchronous remote procedure calls. ParallelX component allows for termination detection and supplies per-process collectives. It

“addresses the challenges of starvation, latency, overhead, waiting, energy and reliability”.

Finally, it supports OpenCL to use distributed GPU and coprocessors. HPX-5 could be compiled on various OS platforms, however it was only tested on several Linux and Darwin (10.11) platforms. Required configurations and environments could be accessed via [407].

3.184 HTCondor



title	HTCondor
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

HTCondor is a specialized workload management system for compute-intensive jobs. HTCondor provides various features like (a) job queuing mechanism, (b) scheduling policy, (c) resource monitoring, (d) priority scheme and (e) resource management just as other full-featured batch systems.

“Users submit their serial or parallel jobs to HTCondor, HTCondor places them into a queue, chooses when and where to run the jobs based upon a policy, carefully monitors their progress, and ultimately informs the user upon completion”.

HTCondor can be used to manage a cluster of dedicated compute nodes. HTCondor uses unique mechanisms to harness wasted CPU power from idle desktop workstations.

“The ClassAd mechanism in HTCondor provides an extremely flexible and expressive framework for matching resource requests (jobs) with resource offers (machines). Jobs can easily state both job requirements and job preferences. HTCondor incorporates many of the emerging Grid and Cloud-based computing methodologies and protocols” [408].

3.185 HTTP



title	HTTP
status	90
section	Data Transport
keywords	Data Transport

3.186 HUBzero



title	HUBzer
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

HUBzero is a collaborative framework which allows creation of dynamic websites for scientific research as well as educational activities. HUBzero lets scientific researchers work together online to develop simulation and modeling tools. These tools can help you connect with powerful Grid computing resources as well as rendering farms [409]. Thus allowing other researchers to access the resulting tools online using a normal web browser and launch simulation runs on the Grid infrastructure without having to download or compile any code. It is a unique framework with simulation and social networking capabilities [410].



title	Hyper-V
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Hypervisors [411] are virtualization platforms that can be used to run virtual environments and manage multiple operating systems on a single server. Hyper-V [412] is Microsoft's hypervisor for the server world. There are 2 different type of hypervisors - type 1 or type 2. While type 1 hypervisors run directly on the system hardware, type 2 need a host operating system to provide management services. Hyper-V is of type-1.

The Hyper-V Server bolsters remote access through Remote Desktop Connection. Organization and arrangement of the host OS and the visitor virtual machines are for the most part done over the system, utilizing either Microsoft Management Consoles on another Windows PC or System Center Virtual Machine Manager. This permits significantly less demanding point and snap design and checking of the Hyper-V Server [413].

3.187.1 Hyper-V Architecture:

It employs isolation of virtual machines through partitions. Partitions are the logical units of isolation where the operating systems complete executions. The high-level architecture of the Hyper-V may comprise 2 types of partitions [414]- * Root partition: This is the parent partition and has direct access to the hardware devices. It has the ability to create child partitions using Hypervisor API. * Child partitions: These partitions do not have access to hardware devices and are present as virtual devices or VDevs. They access the memory and devices by requesting access through the VMBus.

3.187.2 Advantages of Hyper-V over other hypervisors:

1. It's very easy to configure, manage and integrate with other Microsoft products.
2. You don't have to come up with a storage pool when you initially configure it, you can use local storage. All your virtual machines can reside on local storage.
3. It provides scalability and has very innovative features. It favors numerous operating systems. It has a built-in dynamic memory option which will allocate as much memory as the server needs dynamically.
4. Unlike its competitors, you don't need to have separate management software to manage the hypervisor. You can manage your virtual machines right from the server.
5. Unlike type 2 hypervisors, it won't shut down or suspend the machine when you exit out of the window. The virtual machine is always running until you manually exit.



title	IBM BlueMix
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks, retired

Retired technology

BlueMix [415] was IBM's main cloud software prior to its merger with IBM's Cloud brand. Due to this fact, I decided to write this summary on IBM Cloud because it is easier to find better information on this topic, compared to IBM BlueMix. IBM BlueMix was a cloud computing service that was used as a Platform as a Service (PaaS) [416]. According to NetworkWorld, PaaS is used by developers to develop and fix errors without messing with the infrastructure [417]. Therefore, the platform is set up to help developers develop their software without having to completely launch the product and make changes after while the users are already exposed to the software. This helps developers minimize errors and make themselves as ready as possible for launching the product to the market.

With the rebrand to IBM Cloud, they now offer Software as a Service (SaaS) and Infrastructure as a Service (IaaS). With the SaaS addition, developers are able to create their own Blockchain on the cloud, while adding peers to join and edit as well [418]. The SaaS addition provides a unique opportunity to developers to create software together through a smooth interface. Since Blockchain is becoming more and more popular, the addition of Blockchain adaptability is important for developers to continue to use IBM Cloud as their go-to code-writing platform.

With the IaaS addition, users can use services such as compute power, storage, and networking through the cloud [416]. The IaaS

offers a way to store and communicate data cheaper and faster than ever before. The wide variety of services offered through IBM Cloud has made it very successful in the cloud services industry.

Corporations are using IBM Cloud to adapt their operations systems to cut costs and errors to create an efficient supply chain. One example is Walmart, who is partnering with Microsoft's Azure and the IBM Cloud to build their Blockchain and Internet of Things components [419]. IBM Cloud can be used by small developers and giant corporations alike, demonstrating impressive scalability. This is one of the reasons that IBM Cloud has been so successful and a main competitor in the cloud services industry.

3.189 IBM Cloudant



title	IBM Cloudant
status	90
section	NoSQL
keywords	NoSQL

Cloudant is based on both Apache-backed CouchDB project and the open source BigCouch project. IBM Cloudant is an open source non-relational, distributed database service as service (DBaaS) that provides integrated data management, search and analytics engine designed for web applications. Cloudant's distributed service is used the same way as standalone CouchDB, with the added advantage of data being redundantly distributed over multiple machines [420].

3.190 IBM dashDB



title	IBM dashDB
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

IBM dashDB is a data warehousing service hosted in cloud, This aims at integrating the data from various sources into a cloud data base. Since the data base is hosted in cloud it would have the benifits of a cloud like scalability and less maintainance. This data base can be configured as ‘transaction based’ or ‘Analytics based’ depending on the work load [421]. This is available through ibm blue mix cloud platform.

dash DB has build in analytics based on IBM Netezza Analytics in the PureData System for Analytics. Because of the build in analytics and support of in memory optimization promises better performance efficiency. This can be run alone as a standalone or can be connected to variousBI or analytic tools [422].

3.191 IBM Spectrum Scale, formerly GPFS



title	IBM Spectrum Scale, formerly GPFS
status	10
section	File systems
keywords	File systems

General Parallel File System (GPFS) was rebranded as IBM Spectrum Scale on February 17, 2015 [423].

Spectrum Scale is a clustered file system, developed by IBM, designed for high performance. It

"provides concurrent high-speed file access to applications executing on multiple nodes of clusters"

and can be deployed in either shared-nothing or shared disk modes [423]. Spectrum Scale is available on AIX, Linux, Windows Server, and IBM System Cluster 1350 [423]. Due to its focus on performance and scalability, Spectrum Scale has been utilized in compute clusters, big data and analytics - including support for Hadoop Distributed File System (HDFS), backups and restores, and private clouds [424].

3.192 IBM System G



title	IBM System G
status	90
section	Application and Analytics
keywords	Application and Analytics

IBM System G provides a set of Cloud and Graph computing tools and solutions for Big Data [425]. In fact, the G stands for Graph and typically spans a database, visualization, analytics library, middleware and Network Science Analytics tools. It assists the easy creating of graph stores and queries and exploring them via interactive visualizations [426]. Internally, it uses the property graph model for its working. It consists of five individual components - gShell, REST API, Python interface to gShell, Gremlin and a Visualizer. Some of the typical applications wherein it can be used include Expertise Location, Commerce, Recommendation, Watson, Cybersecurity, etc [427].

However, it is to be noted that the current version does not work in a distributed environment and it is planned that future versions would support it.

IBM Watson



title	IBM Watson
status	95
section	TBD
keywords	TBD

3.192.1 Old: IBM Watson

IBM Watson is a super computer built on cognitive technology that

processes information like the way human brain does by understanding the data in a natural language as well as analyzing structured and unstructured data [428]. It was initially developed as a question and answer tool more specifically to answer questions on the quiz show *Jeopardy* but now it has been seen as helping doctors and nurses in the treatment of cancer. It was developed by IBM's DeepQA research team led by David Ferrucci. With Watson you can create bots that can engage in conversation with you [429]. You can even provide personalized recommendations to Watson by understanding a user's personality, tone and emotion. Watson uses the Apache Hadoop framework in order to process the large volume of data needed to generate an answer by creating in-memory datasets used at run-time. Watson's DeepQA UIMA s (Unstructured Information Management Architecture) annotators were deployed as mappers in the Hadoop Map-Reduce framework. Watson is written in multiple programming languages like Java, C++, Prolog and it runs on the SUSE Linux Enterprise Server. Today Watson is available as a set of open source APIs and Software As a Service product as well [429].

3.192.2 New: IBM Watson

IBM's Watson computer, named after the company's first CEO, Thomas Watson, was created to fulfill an engineering challenge: to defeat Jeopardy! champions [430]. Although Watson defeated the champions, that victory was years in the making. Numerous adjustments were required to make all 3 contestants equal. For example, Watson was required to push a button, like the other contestants [430]. Since the early days of Watson, many advances in computing have happened. The explosion of machine learning algorithms and artificial intelligence have greatly expanded the use of Watson. Watson has developed into a marketing slogan, and is even co-branded with companies, such as H&R Block [430]. Advancements in hardware enabled the revolution in data analysis that is currently happening. Super computing, like Watson, along with enhanced algorithms have fueled the revolution. Based on such advances, Watson is now a stand-alone division at IBM [430]. Being a separate division and an individual brand gives IBM the ability to deploy

Watson to a myriad of activities. Watson is the hub of IBM's cloud computing business. Watson also forecasts the weather and conducts research from any number of organizations [430].

3.193 ImageJ



title	ImageJ
status	90
section	Application and Analytics
keywords	Application and Analytics

ImageJ is a Java-based image processing program developed at the National Institutes of Health (NIH). ImageJ was designed with an open architecture that provides extensibility via Java plugins and recordable macros. Using ImageJ's built-in editor and a Java compiler, it has enabled to solve many image processing and analysis problems in scientific research from three-dimensional live-cell imaging to radiological image processing. ImageJ's plugin architecture and built-in development environment has made it a popular platform for teaching image processing [431].

3.194 Impala



title	Impala
status	90
section	High level Programming
keywords	High level Programming

Cloudera Impala is Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop [432]. It allows users to execute low latency SQL queries for data stored in HDFS and HBase, without any movement or transformation of data. The Apache Hive provides a powerful query mechanism for hadoop users, but the query response time are not acceptable due to Hive's reliance on MapReduce. Impala technology by Cloudera has its MPP query engine written in C++ replacing the Java engine proves to improve the interactive Hadoop queries and interactive query response time for hadoop users [433] . Impala is faster than Hive also because it executes the SQL queries natively without translating them into Hadoop MapReduce jobs, thus taking less time. Impala uses HiveQL as programming interface and also the Impala's Query Exec Engines are co-located with the HDFS data nodes, so that the data nodes and processing tasks are co-located, following the haddops paradigm [433]. Impala can aslo use Hbase as a data source. Thus, Impala can be considered as an extension to the Apache Hadoop, providing a better performance alternative to Hive-on-top-of-MapReduce model.

Hive and other frameworks built on MapReduce are best suited for long running batch jobs, such as those involving batch processing of Extract, Transform, and Load (ETL) type jobs [432]. The important applications of Impala are when the data is to be partially analyzed or when the same kind of query is to be processed several times from the dataset. When the data is to be partially analyzed, Impala uses parquet as the file format, which is developed by Twitter and

Cloudera and it stores data in vertical manner [434]. When Parquet queries the dataset it only reads the column split part files rather than reading the entire dataset as compared to Hive.

3.195 Inca



title	Inca
status	10
section	Monitoring
keywords	Monitoring

Inca is a grid monitoring [435] software suite. It provides grid monitoring features. These monitoring features provide operators failure trends, debugging support, email notifications, environmental issues etc. [436]. It enables users to automate the tests which can be executed on a periodic basis. Tests can be added and configured as and when needed. It helps users with different portfolios like system administrators, grid operators, end users etc Inca provides user-level grid monitoring. For each user it stores results as well as allows users to deploy new tests as well as share the results with other users. The incat web ui allows users to view the status of test, manage test and results. The architectural blocks of inca include report repository, agent, data consumers and depot. Reporter is an executable program which is used to collect the data from grid source. Reporters can be written in perl and python. Inca repository is a collection of pre build reporters. These can be accessed using a web url. Inca repository has 150+ reporters available. Reporters are versioned and allow automatic updates. Inca agent does the configuration management. Agent can be managed using the incat web ui. Inca depot provides storage and archival of reports. Depot uses relational database for this purpose. The database is accessed using hibernate backend. Inca web UI or incat provides real time as well as historical view of inca data. All communication between inca components is secured using SSL certificates. It requires user credentials for any access to the system. Credentials are created at the time of the setup and installation. Inca's performance has been phenomenal in production deployments. Some of the deployments are running for more than a decade and has been very stable. Overall Inca provides a solid

monitoring system which not only monitors but also detects problems very early on.

3.196 InCommon



title	InCommon
status	10
section	Monitoring
keywords	Monitoring

The mission of InCommon is to

“create and support a common trust framework for U.S. education and research. This includes trustworthy shared management of access to on-line resources in support of education and research in the United States.” [437]

This mission ultimately is a simplification and an elimination of the need for multiple accounts across various websites that are at risk of data spills or misuse. In the academic setting, this helps assist researchers to focus on their area of study, and enabling the cross collaboration which is happening on a global scale. Currently any two and four year higher education institution that is accredited is eligible for joining InCommon.

3.197 Infinispan



title	Infinispan
status	90
section	In-memory databases/caches
keywords	In-memory databases/caches

Infinispan is a highly available, extremely scalable key/value data store and data grid platform. The design perspective of infinispan is exposing a distributed, highly concurrent data structure to make the most use of modern multi-core as well as multi-processor architectures. It is mostly used as a distributed cache, but also can be used as a object database or NoSQL key/value store [438].

Infinispan is mostly used as a cache store. It is predominantly used for applications that are clustered, and requires a cache coherency for data consistency. Infinispan is written in java and is open source. It is fully transactional. Infinispan is used to add clusterability as well as high availability to frameworks. Infinispan has many use-cases, they are: (1) it can be used as a distributed cache (2) Storage for temporal data, like web sessions, (3) Cross-JVM communication, (4) Shared storage, (5) In-memory data processing and analytics and (6) MapReduce Implementation in the In-Memory Data Grid. It is also used in research and academia as a framework for distribution execution and storage [439].

3.198 iRODS



title	iRODS
status	10
section	File management
keywords	File management

The Integrated Rule-Oriented Data System (iRODS) is open source data management software. iRODS is released as a production-level distribution aimed at deployment in mission critical environments. It virtualizes data storage resources, so users can take control of their data, regardless of where and on what device the data is stored. The development infrastructure supports exhaustive testing on supported platforms. The plugin architecture supports microservices, storage systems, authentication, networking, databases, rule engines, and an extensible API [440]. iRODS implements data virtualization, allowing access to distributed storage assets under a unified namespace, and freeing organizations from getting locked in to single-vendor storage solutions. iRODS enables data discovery using a metadata catalog that describes every file, every directory, and every storage resource in the iRODS Zone. iRODS automates data workflows, with a rule engine that permits any action to be initiated by any trigger on any server or client in the Zone. iRODS enables secure collaboration, so users only need to log in to their home Zone to access data hosted on a remote Zone [441].



o: start with a sentence what
jClouds is

title	JClouds
status	10
section	Interoperability
keywords	Interoperability

use italics and not bold

“Multi-cloud APIs abstract cloud platform differences and provide a single syntax for accessing and managing services from a variety of cloud platforms. Whereas an abstraction layer provides the single syntax, adapters translate service management requests into platform-specific calls to the cloud platform.” ???

jclouds is an open source library for developing multi-cloud applications. Using Java or Clojure languages, developers can develop portable applications on the cloud. It supports 30 cloud providers including Google, Microsoft Azure and Amazon [442]. Thanks to jclouds, developers can migrate the application one cloud to another easily. It also allows users to use cloud-specific features [443].

jclouds provides configurations for Maven, ANT, and Leiningen in the documentation. To use jclouds in a project, it can be included as a dependency [444].

jclouds has 4 main concepts:

1. **Views** are designed to abstract code, so developers can write generic codes which can run on different cloud platforms.

2. **APIs** in JClouds mean calling a specific cloud service to do the job.
3. **Providers** mean a cloud service which provides APIs to use the service.
4. **Context** represents the connection to a provider. Since context creation is expensive, for most cases, creating a single context for a provider and using it until application termination is recommended [445].

jclouds provides several views as Java and Clojure libraries. These are:

ComputeService gives the opportunity to manage instances in the cloud.

Users can start and configure multiple machines at once. It allows to run nodes as sets and manage them by a group name.

The BlobStore API is a portable means of managing key-value storage providers such as Microsoft Azure Blob Service, Amazon S3, or OpenStack Object Storage. It offers a synchronous API to your data." [442]

LoadBalancer API provides an interface to manage supported providers load balancers. It simplifies the load balancer configuration [442].

3.200 Jelastic



title	Jelastic
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Jelastic (acronym for Java Elastic) is an unlimited PaaS and Container based IaaS within a single platform that provides high availability of applications, automatic vertical and horizontal scaling via containerization to software development clients, enterprise businesses, DevOps, System Admins, Developers, OEMs and web hosting providers [446]. Jelastic is a Platform-as-Infrastructure provider of Java and PHP hosting. It has international hosting partners and data centers. The company can add memory, CPU and disk space to meet customer needs. The main competitors of Jelastic are Google App Engine, Amazon Elastic Beanstalk, Heroku, and Cloud Foundry. Jelastic is unique in that it does not have limitations or code change requirements, and it offers automated vertical scaling, application lifecycle management, and availability from multiple hosting providers around the world [447].



title	Jena
status	90
section	NoSQL
keywords	NoSQL, OWL, RDF, Semantic Web, Triples

In 2000 HP Labs began developing Jena and in 2010 Jena became part of the Apache Software Foundation [448]. Jena is open source and written in Java. Jena is a tool for the semantic web with Java libraries and APIs to develop tool and applications. Key distinguishes for Jena is its support of OWL and tools to publish RDF data. The semantic web, sometimes called web 3.0, is based on a concept of storing data differently than conventional relational databases. Storing data in RDF triples allows linkages between data that might otherwise remain disconnected. Jena is an application that provides the tools to help developers build applications to store, query and organize data in this RDF triples format.

There are several key elements to Jena that are best understood with a basic understanding of RDF. RDF or Resource Description Framework is a way to model and connect resources or objects. The building blocks in RDF are triples. Triples are built specifically in the Subject.Predicate.Object format. An example of a triple would be `wine.Color.Red`. The data model that is created storing data this way becomes a semantic graph. Jena is built to provide tools to work with RDF and the semantic web. Jena has some key libraries and capabilities that are key for developers buying into building applications in the ontological model. Jena RDF API is key to reading and building triples and RDF graphs. The Jena RDF API provides the basic create, read, update and delete functionality for dealing with triples in RDF. The Jena ARQ processor is another important tool in Jena's tool kit that enables query using SPARQL `???`. On other key distinguisher of Jena is its support for OWL. OWL or the Web

Ontology Language is way represent and build semantics of data similar but more extensible than RDF [449].

3.202 JGroups



title	JGroups
status	90
section	Monitoring
keywords	Monitoring

3.203 Jitterbit



title	Jitterbit
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Jitterbit is an integration tool that delivers a quick, flexible and simpler approach to design, configure, test, and deploy integration solutions [450]. It delivers powerful, flexible, and easy to use integration solutions that connect modern on premise, cloud, social, and mobile infrastructures. Jitterbit employs high performance parallel processing algorithms to handle large data sets commonly found in ETL initiatives [451]. This allows easy synchronization of disparate computing platforms quickly. The Data Cleansing and Smart Reconstruction tools provides complete reliability in data extraction, transformation and loading.

Jitterbit employs a no-code GUI (graphical user interface) and work with diverse applications such as: ETL (extract-transform-load), SaaS (Software as a Service), SOA (service-oriented architecture).

Thus it provides centralized platform with power to control all data. It supports many document types and protocols: XML, web services, database, LDAP, text, FTP, HTTP (S), Flat and Hierarchic file structures and file shares [452]. It is available for Linux and Windows, and is also offered through Amazon EC2 (Amazon Elastic Compute Cloud). Jitterbit Data Loader for Salesforce is a free data migration tool that enables Salesforce administrators automated import and export of data between flat files, databases and Salesforce.

3.204 JMS



title	JMS
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives

JMS (Java Messaging Service) is a java oriented messaging standard that defines a set of interfaces and semantics which allows applications to send, receive, create, and read messages. It allows the communication between different components of a distributed application to be loosely coupled, reliable, and asynchronous [453]. JMS overcomes the drawbacks of RMI (Remote Method Invocation) where the sender needs to know the method signature of the remote object to invoke it and RPC (Remote Procedure Call), which is tightly coupled i.e it cannot function unless the sender has important information about the receiver.

JMS establishes a standard that provides loosely coupled communication i.e the sender and receiver need not be present at the same time or know anything about each other before initiating the communication. JMS provides two communication domains. A point-to-point messaging domain where there is one producer and one consumer. On generating message, a producer simply pushes the message to a message queue which is known to the consumer. The other communication domain is publish/subscribe model, where one message can have multiple receivers [454].



title	Juju
status	10
section	DevOps
keywords	DevOps

Juju [455] is an open source project developed by Canonical, makers of Ubuntu, with the main purpose of providing a framework that makes it easier to deploy, control, configure, scale, integrate, and monitor services on the cloud [456]. Juju can either be installed directly or be hosted as JAAS, Juju as a Service, and operates on many popular clouds such as Amazon Web Services, Microsoft Azure, Google Compute Engine, OpenStack, VSphere, etc. [457].

Juju's heart is a common controller to manage machines on running application models, also response to system events. Users can interact with Juju's controller via command line or through its GUI interface. Juju controller can manage multiple models and their VMs, access and authorization, and configuration of the models [458].

One of Juju's advantages is its ability to provide solutions called charms or bundles, collection of charms, to help simplify its process and improve user adaptation [459]. Users can download and access charms on charm store community, each charm contains all operations needed for it to run as a single application and can be integrated with other applications. Juju Charm Store lauched on April 3rd, 2012; the store community test their charms very often, allow developers to upload new charms, and allow access for public users to all the charms to increase extensibility and flexibility [457]. Bundles allows the ability to form relationships between charms and their services so that users can deploy multiple charms in one go. Charms and bundles perform many DevOps functions such as installation, configuration, upgrade and update, horizontal server scaling, system

health checks, monitoring, and benchmarking [458].

Juju is often compared or integrated with other alternative DevOps/service orchestration software such as Ansible, Chef, Puppet, etc. [457]. Another advantage of Juju comparing to other software is its dynamic configuration ability which makes it easier to re-configure and modify relationship between services [458].

3.206 Jupyter and IPython



title	Jupyter and IPython
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

The Jupyter Notebook is a language-agnostic HTML notebook web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text [460]. The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results [461]. The Jupyter notebook combines two components:

1. A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
2. Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

Notebooks may be exported to a range of static formats, including HTML (for example, for blog posts), reStructuredText, LaTeX, PDF, and slide shows, via the nbconvert command [462]. Notebook documents contains the inputs and outputs of a interactive session as well as additional text that accompanies the code but is not meant for execution [463]. In this way, notebook files can serve as a complete computational record of a session, interleaving executable code with explanatory text, mathematics, and rich representations of resulting

objects [464]. These documents are internally JSON files and are saved with the .ipynb extension. Since JSON is a plain text format, they can be version-controlled and shared with colleagues [465].



title	Kafka
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Handling large amount of data quickly is a challenge many companies face. Kafka allows companies to do exactly this. Kafka helps in streaming of data input and reading from servers in small streams of data. Kafka architecture involves producers who send message to broker i.e. Kafka server. This data is sent in basic units called messages which is like sending each row from given dataset. Consumers are applications that read data from Kafka servers (brokers). Way Kafka proves efficient is in its ability to handle data fast in real time, compatibility with various languages, scalability, retention abilities among many others. Kafka achieves its fast pace in terms of handling data by breaking the data in smaller sets called partitions. A partition as name suggests is piece of data stored on a machine. Another layer of sophistication that Kafka brings in while handling data is splitting data by topics. This helps when trying to read data from servers. If a consumer wants to read data on certain topic, he needs to know topic, partition and message (also called as offset) number [466].

A Kafka server (broker) can have multiple producers sending data to it simultaneously. At any given time, a Kafka server can have multiple consumers reading data from it. Consumers read the data and either store it on data centers or allow applications to perform real time analytics with it. Consumers can be made efficient by creating multiple consumers that read certain numbers of partitions on given topic. Server itself can be made efficient by dividing the work among cluster of machines. This distributed architecture allows to scale Kafka to handle any size of data. Kafka's ability to support various

languages allows one to create applications as per their expertise [466].

Kafka also helps by giving ability to store data either by date, size or type per users need. Kafka is used by Apple Inc., Netflix, PayPal, Walmart, Uber among many others [467].

3.208 Karajan



o: wrong citation	
title	Karajan
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Karajan is used to allow users to describe various workflows using XML [85]. It also uses a custom yet user friendly language called K. The advantages of using XML and K is that we can use Directed Acyclic Graphs (DAGs) to describe hierarchical workflows. Besides, it is also very easy to handle concurrency using trivial programming constructs like if/while orders. It can also use tools such as Globus GRAM for parallel or distributed execution of various workflows. From an architectural perspective, Karajan mainly consists of three components: Workflow engine, that monitors the execution and is responsible for the higher level interaction with higher level components like the Graphical User Interface Module (GUI) for the description of various workflows; Workflow service, that is used to allow the execution of various workflows using specific functionalities that can be accessed by the workflow engine using specific libraries; and the Checkpointing subsystem that monitors and checks the current state of the workflow. Karajan is typically used as a scientific workflow scheduling technique for various Big Data platforms.

The Karajan code, that can be obtained from Java CoG Kit CVS archive has two interfaces: the command line interface (CLI) and the GUI. The CLI can be accessed via bin/karajan and provides a minimalist interface that is non-interactive and doesn't provide much feedback on the execution status. As against this, the GUI can be accessed via bin/karajan-gui and provides an enriched interface that provides visual features to determine the execution status besides being interactive in real time [468].

3.209 Kepler



title	Kepler
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Kepler, scientific workflow application, is designed to help scientist, analyst, and computer programmer create, execute and share models and analyses across a broad range of scientific and engineering disciplines. Kepler can operate on data stored in a variety of formats, locally and over the internet, and is an effective environment for integrating disparate software components such as merging R scripts with compiled C code, or facilitating remote, distributed execution of models. Using Kepler's GUI, users can simply select and then connect pertinent analytical components and data sources to create a scientific workflow. Overall, the Kepler helps users share and reuse data, workflow, and components developed by the scientific community to address common needs [469].

3.210 Kestrel



title	Kestrel
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Kestrel is a distributed message queue, with added features and bulletproofing, as well as the scalability offered by actors and the Java virtual machine. It supports multiple protocols: memcache: the memcache protocol; thrift: Apache Thrift-based RPC; text: a simple text-based protocol. Each queue is strictly ordered following the FIFO (first in, first out) principle. To keep up with performance items are cached in system memory. Kestrel is more durable as queues are stored in memory for speed, but logged into a journal on disk so that servers can be shutdown or moved without losing any data. When kestrel starts up, it scans the journal folder and creates queues based on any journal files it finds there, to restore state to the way it was when it last shutdown (or was killed or died).

Kestrel uses a pull-based data aggregator system that convey data without prior definition on its destination. So the destination can be defined later on either storage system, like HDFS or NoSQL, or processing system, like storm and spark streaming. Each server handles a set of reliable, ordered message queues. When you put a cluster of these servers together, with no cross communication, and pick a server at random whenever you do a set or get, you end up with a reliable, loosely ordered message queue [470].

3.211 KeystoneML



title	KeystoneML
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

A framework for building and deploying large-scale machine-learning pipelines within Apache Spark. It captures and optimizes the end-to-end large-scale machine learning applications for high-throughput training in a distributed environment with a high-level API [471]. This approach increases ease of use and higher performance over existing systems for large scale learning [471]. It is designed to be a faster and more sophisticated alternative to SparkML, the machine learning framework that is a full member of the Apache Spark club. Whereas SparkML comes with a basic set of operators for processing text and numbers, KeystoneML includes a richer set of operators and algorithms designed specifically for natural language processing, computer vision, and speech processing [471]. It has enriched set of operations for complex domains: vision, NLP, Speech, plus, advanced math And is Integrated with new BDAS technologies: Velox, ml-matrix, soon Planck, TuPAQ and Sample Clean [472].

3.212 Kibana



title	Kibana
status	90
section	Application and Analytics
keywords	Application and Analytics

Kibana is an open source data visualization plugin for Elasticsearch [473]. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster. Users can create bar, line and scatter plots, or pie charts and maps on top of large volumes of data [474]. The combination of Elasticsearch, Logstash, and Kibana (also known as ELK stack or Elastic stack) is available as products or service. Logstash provides an input stream to Elastic for storage and search, and Kibana accesses the data for visualizations such as dashboards [475]. Elasticsearch is a search engine based on Lucene [476]. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Kibana makes it easy to understand large volumes of data. Its simple, browser-based interface enables you to quickly create and share dynamic dashboards that display changes to Elasticsearch queries in real time [477] [478].

3.213 Kite



title	Kite
status	90
section	High level Programming
keywords	High level Programming

Kite is a programming language designed to minimize the required experience level of the programmer. It aims to allow quick development and running time and low CPU and memory usage. Kite was designed with lightweight systems in mind. On OS X Leopard, the main Kite library is only 88KB, with each package in the standard library weighing in at 13-30KB. The main design philosophy is minimalism - only include the minimum necessary, while giving developers the power to write anything that they can write in other languages. Kite combines both object oriented and functional paradigms in the language syntax. One special feature is its use of the pipe character (|) to indicate function calls, as opposed to the period (.) or arrow (->) in other languages. Properties are still de-referenced using the period [479]. Kite also offers a digital assistant for programmers. Kite offers a product which sits as a sidebar in code editor and enables programmers to search for opensource codes to implement in their codes. It even provides relevant examples/syntax and also tries to spot errors in the programs [480].

3.214 KVM



title	KVM
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

It is an acronym for Kernel-based Virtual Machine for the Linux Kernel that turns it into a hypervisor upon installation. It was originally developed by Qumranet in 2007 [481]. It has a kernel model and uses kernel as VMM. It only supports fully virtualized VMs. It is very active for Linux users due to its ease of use, it can be completely controlled by ourselves and there is an ease for migration from or to other platforms. It is built to run on a x86 machine on an Intel processor with virtualization technology extensions (VT-x) or an AMD-V. It supports 32 and 64 bit guests on a 64 bit host and hardware visualization features. The supported guest systems are Solaris, Linux, Windows and BSD Unix [482].

3.215 Kyoto Cabinet



title	Kyoto Cabinet
status	10
section	High level Programming
keywords	High level Programming

Kyoto Cabinet is a library of routines for managing a database which is a simple data file containing records [483]. Each record in the database is a pair of a key and a value. Every key and value is serial bytes with variable length. Both binary data and character string can be used as a key and a value. Each key must be unique within a database. There is neither concept of data tables nor data types. Records are organized in hash table or B+ tree. Kyoto Cabinet runs very fast. The elapsed time to store one million records is 0.9 seconds for hash database, and 1.1 seconds for B+ tree database. Moreover, the size of database is very small. The, overhead for a record is 16 bytes for hash database, and 4 bytes for B+ tree database. Furthermore, scalability of Kyoto Cabinet is great. The database size can be up to 8EB (9.22e18 bytes).

3.216 Kyoto/Tokyo Cabinet



title	Kyoto/Tokyo Cabinet
status	10
section	NoSQL
keywords	NoSQL

Tokyo Cabinet and Kyoto Cabinet are libraries of routines for managing a database [484] [485]. The database normally is a simple data file containing records having a key value pair structure. Every key and value is serial bytes with variable length. Both binary data and character string can be used as a key and a value. There is no concept of data tables nor data types like RDBMS or DBMS. Records are organized in hash table, B+ tree, or fixed-length array. Tokyo and Kyoto cabinets both are developed as a successor of GDBM and QDBM which are library routines for managing database as well. Tokyo Cabinet is written in the C language, and is provided as API of C, Perl, Ruby, Java, and Lua. Tokyo Cabinet is available on platforms which have API conforming to C99 and POSIX. Whereas Kyoto Cabinet is written in the C++ language, and is provided as API of C++, C, Java, Python, Ruby, Perl, and Lua. Kyoto Cabinet is available on platforms which have API conforming to C++03 with the TR1 library extensions. Both are free software licenced under GNU (General Public Licence). Kyoto Cabinet is more powerful and has convenient library structure than Tokyo and recommends people to use Kyoto [484]. Since they use key-value pair concept, you can store a record with a key and a value, delete a record using the key and even retrieve a record using the key. Both have smaller size of database file, faster processing speed and provide effective backup procedures.

3.217 Lambda



title	Lambda
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives

AWS Lambda is a product from amazon which facilitates serverless computing [486]. AWS Lambda allows for running the code without the need for provisioning or managing servers, all server management is taken care by AWS. The code to be run on AWS Lambda is called a server function which can be written in Node.js,Python,Java,C#. Each Lambda function is to be stateless and any persistent data needs are to be handled through storage devices. AWS Lambda function can be setup using the AWS Lambda console where one can setup the function code and specify the event that triggers the functional call. AWS Lamda service supports multiple event sources as identified in [487]. AWS Lambda is designed to use replication and redundancy to provide for high availability both for the service itself and the function it runs. AWS Lambda automatically scales your application by running the code in response to each trigger. The code runs in parallel and processes each trigger individually, scaling precisely with the size of the workload. Billing for AWS Lambda is based on the number of times the code executes and in 100 ms increments of the duration of the processing.

3.218 LDAP



title	LDAP
status	90
section	Monitoring
keywords	Monitoring

LDAP stands for Lightweight Directory Access Protocol. It is a software protocol for enabling anyone to locate organizations, individuals, and other resources such as files and devices in a network, whether on the Internet or on corporate internet. [488]

LDAP is a lightweight (smaller amount of code) version of Directory Access Protocol (DAP), which is part of X.500, a standard for directory services in a network. In a network, a directory tells you where in the network something is located. On TCP/IP networks (including the Internet), the domain name system (DNS) is the directory system used to relate the domain name to a specific network address (a unique location on the network). However, you may not know the domain name. LDAP allows you to search for an individual without knowing where they're located (although additional information will help with the search). An LDAP directory can be distributed among many servers. Each server can have a replicated version of the total directory that is synchronized periodically. An LDAP server is called a Directory System Agent (DSA). An LDAP server that receives a request from a user takes responsibility for the request, passing it to other DSAs as necessary, but ensuring a single coordinated response for the user.

3.219 LevelDB



title	LevelDB
status	10
section	NoSQL
keywords	NoSQL

LevelDB is a light-weight, single-purpose library for persistence with bindings to many platforms [489]. It is a simple open source on-disk key/value data store built by Google, inspired by BigTable and is used in Google Chrome and many other products. It supports arbitrary byte arrays as both keys and values, singular get, put and delete operations, batched put and delete, bi-directional iterators and simple compression using the very fast Snappy algorithm. It is hosted on GitHub under the New BSD License and has been ported to a variety of Unix-based systems, Mac OS X, Windows, and Android. It is not an SQL database and does not support SQL queries. Also, it has no support for indexes. Applications use LevelDB as a library, as it does not provide a server or command-line interface.



title	Libcloud
status	10
section	Interoperability
keywords	Interoperability

Apache project Libcloud [490], originally developed by Cloudkick, is a universal interface between your program and various API's of all popular cloud providers. It was made an open source package from July 2009. Nowadays, there are several different cloud providers in the market with different API's. It has become extremely difficult to communicate with different cloud providers. Libcloud helps users to call different clouds at one place. It also allows you to interact with the cloud server in a simplified way. It sits between the application and all API's of the cloud providers thereby supporting direct integration with the API's. Libcloud is extensively used by Cloud providers enabling multi-cloud integration.

“One Interface to Rule Them All” [491]

Libcloud is a python library which allows users to manage their cloud infrastructure and load balances on the cloud. It supports Python 2.5, 2.6, 2.7, 3 and PyPy [490]. Libcloud also provides deployment functionality which makes it easy to bootstrap a server. Installation, upgrading, and use of Libcloud are easy. It handles XML, JASON and also text-based API. Libcloud provides support to more than 50 providers including Microsoft Azure, Apache Cloud stack, Amazon Web Services, OpenStack and VMware.

Resources that can be managed with Libcloud are divided into the following categories. Cloud Servers and Block Storage such as Amazon's EC2 and Rackspace Cloud Servers, Cloud Object Storage and Content Delivery Network (CDN) such as Amazon's S3 and

Rackspace Cloud Files, Load Balancers as a Service such as Amazons Elastic Load Balancer and Go Grid Load Balancers, Domain Name System (DNS) as a Service such as Amazon's Route 53. Backup as a Service such as Amazon EBS and OpenStack Freezer. It also provides container-based services. Libcloud supports various methods for different APIs [492].

While moving an application to the cloud, it is difficult to choose a single provider. Libcloud has prevented vendor lock-in to rely on a single IT infrastructure provider up to a great extent, reducing dependency and allowing organizations to work with different cloud providers at the same time.

3.221 Libvirt



title	Libvirt
status	10
section	Interoperability
keywords	Interoperability

Libvirt is an open source API to manage hardware virtualization developed by Red Hat. It is a standard C library but has accessibility from other languages such as Python, Perl, Java and others [493]. Multiple virtual machine monitors (VMM) or hypervisors are supported such as KVM,QEMU, Xen, Virtuozzo, VMWare ESX, LXC, and BHyve. It can be divided into five categories such as hypervisor connection, domain, network, storage volume and pool [494]. It is accessible by many operating systems such as Linux, FreeBSD, Mac OS, and Windows OS.

3.222 Ligra



title	Ligra
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Ligra is a Light Weight Graph Processing Framework for the graph manipulation and analysis in shared memory system. It is particularly suited for implementing on parallel graph traversal algorithms where only a subset of the vertices are processed in an iteration. The interface is lightweight as it supplies only a few functions. The Ligra framework has two very simple routines, one for mapping over edges and one for mapping over vertices.

The implementations of several graph algorithms like BFS, breadth-first search, betweenness centrality, graph radii estimation, graph-connectivity, PageRank and Bellman-Ford single-source shortest paths efficient and scalable, and often achieve better running times than ones reported by other graph libraries/systems [495]. Although the shared memory machines cannot be scaled to the same size as distributed memory clusters, but the current commodity single unit servers can easily fit graphs with well over a hundred billion edges in the shared memory systems that are large enough for any of the graphs reported in the paper [496].

3.223 LinkedIn



title	LinkedIn
status	90
section	Streams
keywords	Streams

LinkedIn is a social networking website for Business and employment [497]. LinkedIn has more than 400 million user profiles (as per 10 March 2016 news), and increasing at a rate of 2 new member every second [498]. LinkedIn provides different products like:

- People You May Know
- Skill Endorsements
- Jobs You May Be Interested In
- News Feed Updates

Such products are based on big data. To achieve such big data tasks, LinkedIn has its ecosystem consist of Oracle, Hadoop, Pig, Hive, Azkaban (Workflow), Avro Data, Zookeeper, Aster Data, Data In-Apache Kafka, Data Out- Apache Kafka and Voldemort [498]. LinkedIn uses Hadoop and Aster Data as an analytics layer [499]. LinkedIn partitioned the user's data into separate DB's stored it in XML format. Voldemort is a key lookup system used to store the analytically-derived data for the products like People You May Know. Voldemort stores the data in key-value form [499]. LinkedIn has exposed REST API to get the user data [500].



title	Linux-Vserver
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Linux-VServer is an operating system-level virtualization technology that allows for a number of isolated execution environments to be run within the same operating system kernel [501]. Demands for virtualization in the modern era are high due to the fast-evolving tech markets. The virtualization technology is critical in delivering Infrastructure as a Service (IaaS) [502]. More specifically, high-density cloud server and mobile technologies are in the need of a flexible, low-overhead virtualization [503]. These secluded execution environments are called containers. The basic idea behind the

“kernel subsystem containerization approach is the context isolation” [504],

which according to [504], promotes

“kernel modification to isolate a container into a separate, logical execution context such that it cannot see or impact processes, files, network traffic, global IPC/SHM, etc., belonging to another container” [504].

Moreover, in order to reduce disk space utilization as well as the overall system utilization, Linux-VServer implements a unique filesystem unification model [504]. This model fosters the idea to hard-link the files that are seldomly edited or shared by more than one container on the shared systems [504]. The application of this model is made possible by the guest containers that have the capability to securely share filesystem objects [504]. In general,

virtualization technologies provide aide in many usage scenarios. They

“promise such features as configuration independence, software interoperability, better overall system utilization, resource guarantees” [504],

as well

as “close-to-native performance and density” [504]

and the potential for dynamic resource management improvements [501]. Weaknesses, on the other hand, would include the incapability to handle multiple kernels accessed on multiple operating systems simultaneously [501]. Likewise, the kernel footprint of the Linux-VServer is rather small, however, it is not considered to be fully equipped will all necessary features as its support

“for true network virtualization and container migration” [504]

is still deficient. Overall, container virtualization technologies seem to be very popular and will continue to thrive and move boundaries of the industry away from the virtual machines, despite some limitations. Compared to virtual machines, they are more flexible and incur less costs [505].

3.225 Llama



title	Llama
status	90
section	Cluster Resource Management
keywords	Cluster Resource Management

Llama stands for leveraging learning to automatically manage algorithms. There has been a phenomenal improvement in algorithm portfolio and selection approaches. The main drawback of them is that their implementation is specific to a problem domain and customized which leads to the difficulty of exploring new techniques for certain problem domains. Llama has been developed to provide an extensible toolkit which can initiate exploration of a variety of portfolio techniques over a wide range of problem domains. It is modular and implemented as an R package. It leverages the extensive library of machine learning algorithms and techniques in R [506]. Llama can be regarded as a framework which provides the prerequisites for initiating automatic portfolio selectors. It provides a set of methods for combining several trivial approaches of portfolio selection into sophisticated techniques. The primary reason behind the introduction of Llama was to help the researchers working in algorithm selection, algorithm portfolios, etc. and can be just used as a tool for designing the systems [506].



title	LMDB (key value)
status	10
section	In-memory databases/caches
keywords	In-memory databases/caches

LMDB stands for Lightening Memory-Mapped Database Manager ???, a small software library created by Howard Chu and developed by Symas. Also, the database can be used on almost all modern operating systems. According to Mr. Chu, the database is a very simple key-value store that uses a B+ tree and is fully transactional [507]. LMDB was modeled after the BerkleyDB API, but simplified to cut out unnecessary features. Mr. Chu and his team used BerkleyDB API for many years prior to forming their own software, omitting the BDB features that were found to have no benefit or be problematic [507]. LMDB is essentially a simpler version of the BDB that focuses on the important features, as opposed to the complexity, of the BDB.

LMDB is a memory-mapped database. Therefore, the data that is summoned from the database is found directly through the mapped memory, minimizing user errors and crashes. One of the key points that Mr. Chu pushes is that the system is crash-proof and will never overwrite live data [508]. Therefore, a writer is protected from losing their work during the process of writing the data to the database.

The database uses what Mr. Chu refers to as “concurrency support” to ensure that multi-processes are able to be performed. Only a single writer may have a live transaction at any time. Therefore, writers do not write duplicate or deadlock data. In a presentation at DEVOXX France, Mr. Chu says,

“It is a single-writer model, but multiple readers. So, writers don’t block readers and readers don’t block

writers" [507].

This is a helpful feature for writers and readers alike because they are able to write without fear of a reader blocking their access to the database while trying to write data onto the database. Also, readers are able to read the data even when a writer may be using the database.

3.227 Logstash



title	Logstash
status	90
section	Application and Analytics
keywords	Application and Analytics

Logstash is an open source data collection engine with real-time pipelining capabilities. Logstash can dynamically unify data from disparate sources and normalize the data into destinations of your choice [509]. Cleanse and democratize all your data for diverse advanced downstream analytics and visualization use cases.

While Logstash originally drove innovation in log collection, its capabilities extend well beyond that use case. Any type of event can be enriched and transformed with a broad array of input, filter, and output plugins, with many native codecs further simplifying the ingestion process. Logstash accelerates your insights by harnessing a greater volume and variety of data.



title	Lucene
status	10
section	NoSQL
keywords	NoSQL

Lucene is a widely used text search engine library developed by Apache Software Foundation. It is one of the best tools useful in applications where text search is required. The main feature of Lucene is that it is not only scalable but also performs indexing of documents and querying very fast and efficiently. It can be deployed with a simple API call and can easily be integrated with any applications that are programmed in high level programming languages like Python, Java etc. Hence, this makes it a popular choice for cross platform solutions in text search applications [510]. The documents on which search is performed are simply a collection of fields, which are in turn nothing but a set of field values, which can be numerical or textual data, indexed with a field name. The original text is parsed and converted into a series of tokens, which are stored in fields. In simple words what Lucene basically does is it ranks the existing documents based on the search query and the document with highest rank or relevancy score are retrieved as a search result. The indexing of documents is a dynamic process where Lucene continuously updates the indices of documents as we add new documents and delete or alter the existing ones. Lucene does the search based on terms which are a combination of field name and tokens. Thus, Lucene creates a mapping between terms and documents, which is called Inverted Index, also known as Lucene Index. This mapping facilitates the search process. When there is a search query, the inverted index scores the search results and the document with highest maps referenced to is retrieved as search result. Some of the advantages of Lucene are open source availability, scaling, speed, accuracy, memory efficiency, dynamic and easy to use

and deploy [511].

3.229 Lumberyard



title	Lumberyard
status	90
section	High level Programming
keywords	High level Programming

It is powerful and full-featured enough to develop triple-A, current-gen console games and is deeply integrated with AWS and Twitch (an online steaming service) [512]. Lumberyard's core engine technology is based on Crytek's CryEngine [513]. The goal is

“creating experiences that embrace the notion of a player, broadcaster, and viewer all joining together” [512].

Monetization for Lumberyard will come strictly through the use of Amazon Web Services' cloud computing. If you use the engine for your game, you're permitted to roll your own server tech, but if you're using a third-party provider, it has to be Amazon [514].



title	Lustre
status	10
section	File systems
keywords	File systems

Lustre [515], which is basically a fusion of Linux and cluster is an open-source parallel distributed clustered file system that is designed specifically for high-scalability, high-performance and high-availability [516]. It supports high-performance computing on numerous computer clusters and huge number of nodes. It initially began as a scholarly research venture and was later acquired by Sun Microsystems which has today grown into a file system that supports some of the world's most demanding high-performance computing requirements. Lustre is primarily used by the majority of the high-performance computers available today mainly because of its open-licensing and high-performance capabilities. These benefits make lustre file systems a popular choice for businesses particularly dealing with huge volumes of data spanning across different industries such as meteorology, oil and gas, life sciences etc [517].

Lustre's scalable architecture has three main components - the Metadata Server (MDS), Object Storage Server (OSS), and the clients. This architecture supports multiple computer clusters with thousands of nodes. There are mainly three types of nodes, Client-side nodes, MDS (Meta Data Server) and Object storage servers where we stripe the data all over. All three types of nodes have their own Lustre logic that they have to run but they all converge to the same point when it comes down to networking. Lustre is request response based and all three nodes converge to the same portal RPC (ptlrc) layer. The portal RPC layer does all the processing and passes the requests down to the Lustre networking layer (LNet) [518]. The LNet block represents the common block of code that is done for all Lustre networking. The

LND layer or the LNet Network Driver layer is the code that is specific only to the request that is momentarily being used. The LND layer responds directly to the last layer which is the OFED driver to perform all the infiniband communication. Lustre runs on most commodity hardware and is compatible with any storage device.



title	LXC
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

3.231.1 Old text

LXC (Linux Containers) [ref missing] is an operating-system-level virtualization method for running multiple isolated Linux systems (containers) on a control host using a single Linux kernel [519]. LXC are similar to the traditional virtual machines but instead of having separate kernel process for the guest operating system being run, containers would share the kernel process with the host operating system. This is made possible with the implementation of namespaces and cgroups. [520]

Containers are light weighed (As guest operating system loading and booting is eliminated) and more customizable compared to VM technologies. The basis for docker development is also LXC. [521]. Linux containers would work on the major distributions of linux this would not work on Microsoft Windows.

3.231.2 New text

I propose to improve grammar

LXC is a userspace interface to put different Linux systems in one box. Containers could allow us to manage Linux easily than before due to the complexity of creating and managing the Linux system. LXC provides a powerful tool to make the process convenient for developers. LXC was in active development since 2008, it is famous for the developer who uses Linux containers. Time proofs that LXC

works well in several complex environments. The developers who develop LXC also contributed a lot to the adaption of LXC to Linux kernel, which benefits those developers a lot.

It is a free software virtualization system that LXC provided to a machine that works on GNU/Linux, which could be efficient when dealing with multi-systems. Due to the effectiveness of LXC, developers are able to run several virtual systems at the same time and could manage them in order. The units same as chroots, are able to operate and manage resources by themselves even though they work on the same kernel.

"Current LXC uses the following kernel features to contain processes: Kernel namespaces (ipc, uts, mount, pid, network and user), Apparmor and SELinux profiles, Seccomp policies, Chroots (using pivot_root), Kernel capabilities, CGroups (control groups)" [522]

One thing that needed to point out is that Linux Containers are seemed like kind of chroot or VM, but the LXC is focused on how to establish a place for the most normal Linux installation and share the same kernel with the original one.

Different from the common view of LXC, LXC provides a nearly totally complete operating system-level virtualization which is separated with another environment.

"LXC relies on the Linux kernel groups functionality that was released in version 2.6.24. It also relies on other kinds of namespace isolation functionality, which were developed and integrated into the mainline Linux kernel" [523].

Containers, which are not as our generally thinking, are a different concept with virtual machines, they two have differences in many aspects. Both of them have their own benefits and problems in the practice. VM could establish a separated environment for stuff due to it is a totally new system for the processes. The flexibility allows us to

do different works in the same laptop but with different operating systems, the multi-systems could boost our efficiency in many aspects.

In contrast, containers, they just occupy part of the host's resources and manage them for the tasks. They are not totally separated with the host since they share part of the kernel. With no doubt that containers is good enough to finish many works even though it is not thoroughly isolated.



title	LXD
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

LXD (pronounced lex-dee), Linux Container Hypervisor, is a daemon that acts as the manager for LXC containers, moreover it offers a command line Interface to LXC with help of its Representational State Transfer (REST) API [524]. LXC is a virtualization method for executing several Linux containers on one host machine, simultaneously, utilizing a single kernel [525].

“It was build with aim of providing VM like virtualization with container like performance” [526].

LXD can't be used by itself and is made specifically to bolster the existing LXC features. LXD further adds new functionalities to the containers when plugged to LXC. Firstly, with help of its template distribution system it expedites the container creation process. Second, LXD acts as the LXC container hypervisor and administers its resources like storage, CPU, memory etc making it more scalable. Additionally, it improves the security by introducing resource restrictions. Also, it provides the feature of taking snapshots of the running containers by capturing the container's state and runtime details such that the admin can return to a state anytime if needed. LXD also helps with the live migration of running containers from one server to another without any disruption in the container [525].

LXD uses image-based workflow and therefore the containers are to be created from an image. It supports two types of image formats : unified tarball, split model. In the first format, produced by LXD, the container rootfs i.e., the file systems, and their metadata are stored in

a single tarball. In the latter, the rootfs and metadata are split into two tarballs, this is for easy image building from non-LXD rootfs [525].

Written in Go, LXD is a free software that can be easily downloaded on any system. Canonical first released it in 2014 as an opensource project and it's now available in two kinds of releases: LTS releases and Feature releases, of which LTS is normally used for production environments as it's not regularly updated with new features [526].

3.233 Mahout



title	Mahout
status	10
section	Application and Analytics
keywords	Application and Analytics

“Apache Mahout software provides three major features: (1) A simple and extensible programming environment and framework for building scalable algorithms (2) A wide variety of premade algorithms for Scala Apache Spark, H2O, Apache Flink (3) Samsara, a vector math experimentation environment with R-like syntax which works at scale” [527].

3.234 MapGraph



title	MapGraph
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

3.235 Marionette Collective



title	Marionette Collective
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

It is basically a framework for management of a system where the systems undergo an organized coordination resulting in an automated deployment of systems which creates an orderly workflow or a parallel wise job execution. It doesn't rely on central inventories such as SSH and uses tools such as Middleware [528]. This gives an advantage of delivering a very scalable and quick execution environment. Mcollective gives us a huge advantage of working with a large number of servers, it uses publish/subscribe middleware for communicating with many hosts at once in a parallel manner. Mcollective allows us to interact with a cluster of servers at the same time, it allows us to use a simple command line to call remote agents and there isn't a centralized inventory. Mcollective uses a broadcast paradigm to distribute the requests, where all the servers receives the request at the same time which are also attached with a filter. The servers which match the filter will act on these requests.

3.236 Medusa-GPU



title	Medusa-GPU
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Graphs are commonly used data structures. However, developers may find it challenging to write correct and efficient programs. Furthermore, graph processing is further complicated by irregularities of graph structures. Medusa enables the developers to write sequential C/C++ code. Medusa provides a set of APIs, which embraces a runtime system to automatically execute those APIs in parallel [529]. A number of optimization techniques are implemented to improvise the efficiency of graph processing. Experimental results demonstrate that (1) Medusa greatly simplifies implementation of GPGPU programs for graph processing, with many fewer lines of source code written by developers; (2) The optimization techniques significantly improve the performance of the runtime system, making its performance comparable with or better than manually tuned GPU graph operations [529]. Medusa has proved to be a powerful framework for networked digital audio and video framework and by exploiting the APIs it takes a modular approach to construct complex graph systems [530].

3.237 Megastore and Spanner



title	Megastore and Spanner
status	10
section	NoSQL
keywords	NoSQL

Spanner is Google's distributed database which is used for managing all google services like play, gmail, photos, picasa, app engine etc. Spanner is distributed database which spans across multiple clusters, datacenters and geo locations [531]. Spanner is structured in such a way so as to provide non blocking reads, lock free transactions and atomic schema modification. This is unlike other noSql databases which follow the CAP theory i.e. you can choose any two of the three: Consistency, Availability and Partition-tolerance. However, spanner gives an edge by satisfying all three of these. It gives you atomicity and consistency along with availability, partition tolerance and synchronized replication. Megastore bridges the gaps found in google's bigtable. As google realized that it is difficult to use bigtable where the application requires constantly changing schema. Megastore offers a solution in terms of semi-relational data model. Megastore also provides a transactional database which can scale unlike relational data stores and synchronous replication [532]. Replication in megastore is supported using Paxos. Megastore also provides versioning. However, megastore has a poor write performance and lack of a SQL like query language. Spanners basically adds what was missing in Bigtable and megastore. As a global distributed database spanner provides replication and globally consistent reads and writes. Spanner deployment is called universe which is a collections of zones. These zones are managed by singleton universe master and placement driver. Replication in spanner is supported by Paxos state machine. Spanner was put into evaluation in early 2011 as F1 backend (F1 is Google's advertisement system) which was replacement to mysql. Overall spanner fulfills the needs of

relational database along with scaling of noSQL database. All these features make google run all their apps seamlessly on spanner infrastructure.



title	Memcached
status	10
section	In-memory databases/caches
keywords	In-memory databases/caches

Memcached, developed in 2003, is an open-source memory caching software system that allows for an increase in speed of web-based applications. This tool is helpful to many users and businesses because it minimizes the need for referencing other APIs or external databases. Memcached consists of four major components that processes the ability for storing and retrieving data. Specifically, these include server software, a client-based hashing algorithm, client software, and a line-replaceable unit.

With Memcached, clients can request data if it is stored in the cache and it will automatically return the requested data. However, if the request is not stored within the cache, Memcached will search the database to find it and then store it in Memcached. This data can be referenced in memory along key-values. Memcached includes its own API that is written in multiple languages. If information is changed or expired, Memcached so that the newest content is being delivered. Values that data is stored in is kept in RAM. If RAM runs out, the longest-held data is disposed of. It is not necessary that users use their RAM for Memcached; many people use machines that are specifically for Memcached use only. Data is also only sent to one server upon which data cannot be shared [533].

Each server that utilizes Memcached accesses data through one combined cache and one pool of information. Memcached multiplies the amount of storage a user can obtain when it is logically combined with one or more servers.



title	Mesos
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

Apache Mesos is an open source computational cluster management system that handles distributed environment through dynamic resources allocation, sharing, management and isolation. It can handle very large scale distributed data systems by creating clusters and parallel processing mechanism on the existing resources or computing systems by acting as a resource management system. Apache Mesos is an alternative to Hadoop because in Hadoop, static partition of nodes is done, this is not efficient and can affect system performance largely. Mesos can handle these problems with Fine grained sharing and Two-level sharing. Fine grained sharing allows to make framework that allows to achieve data locality by reading data stored on each machine [534]. The Mesos Master is central to the cluster that identifies if the cluster is available. It has the primary user interface which provides complete information about the clusters that are available. Mesos Agent is yet another important component that holds and manages containers and manages the communication between the local and master. It manages and updates the status of current running tasks to the scheduler. The Mesos Framework is another component that has two parts -

- The Scheduler
- The Executer

The scheduler is within the Mesos master and it schedules the tasks based on the available resources and requirements for the task to complete. The Executer executes the tasks scheduled by the Scheduler and reports the status of each task to the Scheduler and

Mesos manager. One more interesting thing about Apache Mesos is that it handles both memory and CPU scheduling unlike YARN which can handle only memory scheduling. The other features of it are Web based user interface to monitor the state of cluster, resource scheduling like memory, CPU, and it can share the resources across various frameworks. Major users of Apache Mesos for large scale data management infrastructure are Twitter, Airbnb, Xogito [535].



title	MLbase
status	90
section	Application and Analytics
keywords	Application and Analytics

MLbase is a tool designed to simplify the process of testing and selecting suitable machine learning algorithms for a given dataset [536]. The major problem that the tool aims to address is the time and expertise needed in order to learn something about a particular dataset. The user still needs to have some knowledge of machine learning algorithms but MLbase makes some of the processes more accessible to someone who may not be an expert in the area. As an alternative to learning the intricate details of machine learning, implementing, and validating multiple algorithms MLbase provides tools to test and identify the most effective algorithms for a particular dataset [537]. To do this MLbase provides a declarative language which allows the user to optimize and find the best machine learning algorithms for a chosen machine learning task [536].

The MLbase architecture is a distributed architecture in which the master node reads the declared request and creates a “logical learning plan” which determines how the workload is spread across the nodes [536]. These distributed worker nodes are implemented utilizing Apache Spark. As a part of the design the optimization process aims to reduce the search space so that the testing of various machine learning algorithms can be completed in a reasonable time frame. After the logical learning plan is created it is converted into a “physical learning plan” to be executed [536]. The result of this process is a model that the end user can use to make sense of their data. The MLbase query also provides a summary of the quality and the learning process of the models that were returned [536].

Currently, MLbase is still in the early stages of development but tools such as MLbase will become increasingly important as businesses, scientists and others look to make sense of their data. Similar tools exist such as Weka, MADLib, and Mahout but MLbase goes a step further and addresses the problem of algorithm optimization [536]. The MLbase team continues to develop the tool and is working on improving the optimization algorithm, adding advanced machine learning algorithms to the capability of the tool and also visualization tools [537].

3.241 MLlib



title	MLlib
status	90
section	Application and Analytics
keywords	Application and Analytics

MLlib is Apache Spark's scalable machine learning library [538]. Its goal is to make machine learning scalable and easy. MLlib provides various tools such as, algorithms, feature extraction, utilities for data handling and tools for constructing, evaluating, and tuning machine learning pipelines. MLlib uses the linear algebra package Breeze, which depends on netlib-java for optimized numerical processing. MLlib is shipped with Spark and supports several languages which provides functionality for wide range of learning settings. MLlib library includes Java, Scala and Python APIs and is released as a part of Spark project under the Apache 2.0 license [539].

3.242 mlp y fa18-523-68



title	mlpy
status	10
section	Application and Analytics
keywords	Application and Analytics

mlpy is an open source python library made for providing machine learning functionality. It is built on top of popular existing python libraries of NumPy, SciPy and GNU scientific libraries (GSL). It also makes extensive use of Cython language. These form the prerequisites for mlpy. [540] explains the significance of its components: NumPy, SciPy provide sophisticated N-dimensional arrays, linear algebra functionality and a variety of learning methods, GSL, which is written in C, provides complex numerical calculation functionality.

mlpy provides a wide range of machine learning methods for both supervised and unsupervised learning problems. mlpy is multiplatform and works both on Python 2 and 3 and is distributed under GPL3. Mlpy provides both classic and new learning algorithms for classification, regression and dimensionality reduction. A detailed list of functionality is offered by mlpy [541]. Though developed for general machine learning applications, mlpy has special applications in computational biology, particularly in functional genomics modeling.

3.243 Moab



title	Moab
status	90
section	Cluster Resource Management
keywords	Cluster Resource Management

Moab HPC Suite is a workload management and resource orchestration platform that automates the scheduling, managing, monitoring, and reporting of HPC workloads on massive scale. It uses multi-dimensional policies and advanced future modeling to optimize workload start and run times on diverse resources. It integrates and accelerates the workloads management across independent clusters by adding grid-optimized job submission. Moab's unique intelligent and predictive capabilities evaluate the impact of future orchestration decisions across diverse workload domains (HPC, HTC, Big Data, and Cloud VMs) [542].

3.244 MongoDB



title	MongoDB
status	90
section	NoSQL
keywords	NoSQL

MongoDB is a NoSQL database which uses collections and documents to store data as opposed to the relational database where data is stored in tables and rows. In MongoDB a collection is a container for documents, whereas a document contains key-value pairs for storing data. As MongoDB is a NoSQL database, it supports dynamic schema design allowing documents to have different fields. The database uses a document storage and data interchange format called BSON, which provides a binary representation of JSON-like documents.

MongoDB provides high data availability by way of replication and sharding. High cost involved in data replication can be reduced by horizontal data scaling by way of shards where data is scattered across multiple servers. It reduces query cost as the query load is distributed across servers. This means that both read and write performance can be increased by adding more shards to a cluster. Which document resides on which shard is determined by the shard key of each collection.

As far as data backup and restore is concerned the default MongoDB storage engines natively support backup of complete data. For incremental backups one can use MongoRocks that is a third party tool developed by Facebook.

3.245 MQTT



title	MQTT
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Message Queueing Telemetry Transport (MQTT) protocol is an Interprocess communication protocol that could serve as better alternative to HTTP in certain cases [543]. It is based on a publish-subscribe messaging pattern. Any sensor or remote machine can publish its data and any registered client can subscribe the data. A broker takes care of the message being published by the remote machine and updates the subscriber in case of new message from the remote machine. The data is sent in binary format which makes it use less bandwidth. It is designed mainly to cater to the needs to devices that has access to minimal network bandwidth and device resources without affecting reliability and quality assurance of delivery. MQTT protocol has been in use since 1999. One of the notable work is project Floodnet, which monitors river and floodplains through a set of sensors [544].

3.246 MR-MPI



title	MR-MPI
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

MR-MPI stands for Map Reduce-Message Passing Interface is open source library build on top of standard MPI [545]. It basically implements mapReduce operation providing a interface for user to simplify writing mapReduce program. It is written in C++ and needs to be linked to MPI library in order to make the basic map reduce functionality to be executed in parallel on distributed memory architecture. It provides interface for c, c++ and python. Using C interface the library can also be called from Fortran.



title	MRQL
status	10
section	High level Programming
keywords	High level Programming

MRQL (MapReduce Query Language) is a Query processing system which also provides simple data analytics support through SQL-like queries. Apache Hive, Impala and drill provide basic SQL-like functionalities for querying data stored in a distributed environment such as Apache Hadoop and Hama, but when it comes to the application of advanced data analytics algorithms to derive insight from the data, it gets quite complicated with them. So, with MRQL, the implementation of complex Machine learning algorithms used to perform tasks such as Clustering, or indexing algorithms such as the PageRank algorithm etc. on the data present in the HDFS system can be done with ease. While the default mode of operation of MRQL is the MapReduce mode, it can also be used in the Bulk Synchronous Parallel (BSP) mode [546].

“With the BSP mode, it achieves lower latency and higher speed” [546].

In a test performed to compare the BSP mode with the MapReduce in performing K-means clustering, it was found that the BSP mode was faster than the MapReduce mode by an order of magnitude 3. There are a couple of other modes of operation as well - Spark mode (Using Apache Spark) and Flink mode (Using Apache Flink) [547]. MRQL flexibility lies in the fact that it can perform data analysis over diverse data formats such as XML, JSON, Binary and CSV, without the use of any complex MapReduce code [546]. The simplicity of MRQL as compared to other MapReduce based query languages is exemplified by the fact that MRQL allows nested queries while the latter

languages implement the functionality of nested queries by a combination of group-bys and outer joins. Not only is it easier to implement, the elimination of outer-joins lets the optimizer to use the optimal evaluation method. MRQL is extremely versatile as well, allowing the creation of User Defined Functions, User Defined Aggregations and User-Defined Parsers [548].



title	MySQL
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

MYSQL is one of the most widely used systems to manage relational databases, which hold collections of data [549]. The data in those collections is structured in tables, which are made up of rows and columns [549]. The basic idea behind this technology is not only the housing and security of the data, but also to allow other tasks such as data querying, analysis, sorting, and processing to be performed [549]. Another important component of MySQL are the stored procedures that are essentially functions to perform certain tasks that include modifying, updating, inserting, and displaying data. All tasks can be performed on users' local machines as well as on networks, which means that MYSQL can be considered to be a database server [549]. The communication between the client and the server is completed with the use of the standardized Structured Query Language (SQL) [549]. Commands written in this language are used to add new data to the database or maintain/administer the existing data in the database. Depending on which platform the user is utilizing, the SQL language may vary in regard to the dialect; however, the basic commands such as SELECT remain standard across all forms of the language. In today's business environment, this technology is usually used as a back-end application to support their front-end applications user interface, due to the language's stability and speed. Other than these benefits, MYSQL is also very easy to learn because the language is very intuitive. It can also be easily implemented on most of the popular operating systems such as Windows, Linux, Mac OS X, and Unix [549]. Even though these benefits have made MYSQL an industry standard, this technology still has some technical limitations. According to some developers,

| “it is not easy to create incremental back-ups” [550],

and

| “there is no support for XML and OLAP” [550].

Moreover, in some cases where the amount of data increases significantly, the query speed may be noticeably compromised.

3.249 N1QL



title	N1QL
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

3.250 Nagios



title	Nagios
status	90
section	Monitoring
keywords	Monitoring

Nagios [551] is a platform, which provides a set of software for network infrastructure monitoring. It also offers administrative tools to diagnose when failure events happen, and to notify operators when hardware issues are detected. Specifically, illustrates that Nagios is consist of modules including [552]: a core and its dedicated tool for core configuration, extensible plugins and its frontend. Nagios core is designed with scalability in mind. Nagios contains a specification language allowing for building an extensible monitoring systems. Through the Nagios API components can integrate with the Nagios core services. Plugins can be developed via static languages like C or script languages. This mechanism empowers Nagios to monitor a large set of various scenarios yet being very flexible. [553] Besides its open source components, Nagios also has commercial products to serve needing clients.

3.251 Naiad



title	Naiad
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Naiad is a distributed system based on computational model called Timely Dataflow developed for execution of data-parallel, cyclic dataflow programs [554]. It provides an in-memory distributed dataflow framework which exposes control over data partitioning and enables features like the high throughput of batch processors, the low latency of stream processors, and the ability to perform iterative and incremental computations. The Naiad architecture consists of two main components: (1) incremental processing of incoming updates and (2) low-latency real-time querying of the application state.

Compared to other systems supporting loops or streaming computation, Naiad provides support for the combination of the two, nesting loops inside streaming contexts and indeed other loops, while maintaining a clean separation between the many reasons new records may flow through the computation [555].

This model enriches dataflow computation with timestamps that represent logical points in the computation and provide the basis for an efficient, lightweight coordination mechanism. All the above capabilities in one package allows development of High-level programming models on Naiad which can perform tasks as streaming data analysis, iterative machine learning, and interactive graph mining. On the contrary, it is a public reusable low-level programming abstractions leads Naiad to outperforms many other data parallel systems that enforce a single high-level programming model.

3.252 NaradaBrokering



title	NaradaBrokering
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

NaradaBrokering is a content distribution infrastructure for voluminous data streams [556]. The substrate places no limits on the size, rate and scope of the information encapsulated within these streams or on the number of entities within the system. The smallest unit of this substrate called as broker, intelligently process and route messages, while working with multiple underlying communication protocols. The major capabilities of NaradaBrokering consists of providing a message oriented middleware (MoM) which facilitates communications between entities (which includes clients, resources, services and proxies thereto) through the exchange of messages and providing a notification framework by efficiently routing messages from the originators to only the registered consumers of the message in question [557]. Also, it provides salient stream oriented features such as their Secure end-to-end delivery, Robust disseminations, jitter reductions.

NaradaBrokering incorporates support for several communication protocol such as TCP, UDP, Multicast, HTTP, SSL, IPSec and Parallel TCP as well as supports enterprise messaging standards such as the Java Message Service, and a slew of Web Service specifications such as SOAP, WS-Eventing, WS-Reliable Messaging and WS-Reliability [558].

3.253 Neo4J



title	Neo4J
status	10
section	NoSQL
keywords	NoSQL

Neo4J is a popular ACID compliant graph database management system developed by Neo technology [559]. In this database everything is stored as nodes or edges, both of which can be labeled. Labels help in narrowing and simplifying the search process through the database [560]. It is a highly scalable software and can be distributed across multiple machines. The graph query language that accompanies the software has traversal framework which makes it fast and powerful [561]. The Neo4J is often used for clustering. It offers two feature clustering solutions: Causal Clustering and Highly available clustering [562]. Casual clustering focuses on safety, scalability and causal consistency in the graph [563]. The highly available cluster places importance to fault tolerance as each instance in the cluster has full copies of data in their local database.



title	Neptune
status	10
section	Streams
keywords	Streams

Neptune [564] is effective big data technologies for executing vast and complex technological tasks with remarkable efficiency. Nearly every sphere of life is largely dependent on the efficiency of big data technology for solutions that cannot be provided by alternative technological systems. For example, big data technology has often been utilized to improve the efficiency of marketing campaigns by identifying user web clicks, which are used for determining the current shopping trends among the consumers ??? . Normally, such a process would require extensive and intensive calculations using vast amounts of data from databases and other sites.

Among the distinguishing merits of this technology is the capacity to handle large volume of data at exceptionally high velocities. Moreover, the advantage of Neptune is that it is designed with the three advantages, namely greater velocity, variety, and simplicity of operation. This means that it is capable of completing difficult tasks a little bit faster in comparison to ordinary systems that will require more time to perform the same.

3.255 NetCDF



title	NetCDF
status	90
section	File management
keywords	File management

NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array oriented scientific data. NetCDF was developed and is maintained at Unidata, part of the University Corporation for Atmospheric Research (UCAR) Community Programs (UCP). Unidata is funded primarily by the National Science Foundation [565] [566]. The purpose of the Network Common Data Form (netCDF) interface is to support the creation, efficient access, and sharing of data in a form that is self-describing, portable, compact, extendible, and archivable. Version 3 of netCDF is widely used in atmospheric and ocean sciences due to its simplicity. NetCDF version 4 has been designed to address limitations of netCDF version 3 while preserving useful forms of compatibility with existing application software and data archives [565]. NetCDF consists of: (a) A conceptual data model (b) A set of binary data formats (c) A set of APIs for C/Fortran/Java

3.256 Netty



title	Netty
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Netty

“is an asynchronous event-driven network application framework for rapid development of maintainable high performance protocol servers and clients” [567].

Netty

“is more than a collection of interfaces and classes; it also defines an architectural model and a rich set of design patterns” [568].

It is protocol agnostic, supports both connection oriented protocols using TCP and connection less protocols built using UDP. Netty offers performance superior to standard Java NIO API thanks to optimized resource management, pooling and reuse and low memory copying.



title	NiFi (NSA)
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration, ETL, Apache, Open Source,

NiFi is a customizable tool for building flexible data flows while preserving data provenance and security [569]. NiFi provides the ability to build or alter an ETL flow with a few clicks. NiFi builds Gets, Converts, and Pulls in a GUI and allows the user to build and customize the flow [570]. This flexibility and usability is key to NiFis value in a big data world where stovepipes and inflexibility are frequently challenges.

In the world of big data ETL, or Extract Transform and Load, is a prevalent in most big date projects or architecture. If the data being used is in the perfect format and structure, and the data is housed or collected in the ideal location for the end use of the data, then ETL may be superfluous. Otherwise Extract Transform and Load concepts will come into play. However, enabling ETL is frequently more difficult than it sounds. Data moving between systems effectively is tricky to setup and challenging to refactor on the fly when conditions change. NiFi was first developed at the National Security Agency but was released as open source project to the public.

"NiFi was submitted to The Apache Software Foundation (ASF) in November 2014 as part of the NSA Technology Transfer Program" [571].

Since then, Apache Foundation has used it's volunteer organization to grow and mature the project [570]. NiFi incorporates a straightforward UI to engineer traceable data provence with configurable components. NiFi offers up the ability to custom build

processors and incorporate them into a highly customizable flows. Through

“... data routing, transformation, and system mediation logic” [569]

NiFi seeks to automate data flows in a big data environment and give architects the ability to keep data flowing between evolving systems quickly. Amongst a host of features NiFi offers, one sticks out as particularly important because of the challenges associated with what the feature addresses: data errors, data inconsistency, and data irregularity handling. NiFi provides users the ability to incorporate in the flow processes to catch these non-happy path realities in big data. As new situations are discovered a user can quickly build if-then forks in the process to catch, store, or resolve the data issues.



title	Nimbus
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Cloud computing has 3 models; IaaS, PaaS and SaaS. IaaS model is chosen when one wants access to technologies and resources like traditional data centers without needing investment in planning, actual maintenance and management of them. In IaaS model, vendor is responsible for providing data storage, servers, processing power and virtualization while user is responsible for data that they want to handle, applications that provide interaction with data and cloud servers etc. This way users can build virtual data centers. Nimbus is one such private open source could computing IaaS platform. Nimbus has solutions geared towards needs of services to scientific community [572].

Nimbus has two platforms, Nimbus Platform and Nimbus Infrastructure. Nimbus Platform provides set of tools that provide IaaS' scalability and flexibility. Nimbus Infrastructure provides IaaS features that are useful for scientific community. This allows users to utilize virtual machines on available resources to design an infrastructure they want for their purpose. For scientific community, the computations they need to perform are time consuming and complex and need the job divided on cloud to get it done within desired time. However, when they don't have control over architecture of cloud model, their job can get scheduled and be in line without knowing when the task will be done. Nimbus allows user to create multiple virtual machines to be deployed throughout cloud to spread the workload. User can also configure these virtual machines so that they work in tandem and complete their individual tasks to get whole task done in most time and resource efficient manner. This

way users have control over how and when their computational tasks are completed [572].

Nimbus provides speed by providing readymade execution templates and running efficient tasks in C++ [573]. One of many famous use cases for Nimbus is STAR nuclear physics experiment. Nimbus was used to perform complex and otherwise time-consuming computations by using virtual machines across cloud [574].

3.259 (Ninefold)



title	(Ninefold)
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

The Australian based cloud computing platform has shut down their services since January 30, 2016 [575].



title	NWB
status	10
section	Application and Analytics
keywords	Application and Analytics

NWB (The Network Workbench) is a network analysis tool used by professionals of diverse research fields for modelling and visualization of complex datasets [576].

"It is built on Cyberinfrastructure Shell (CIShell) (Cyberinfrastructure for Network Science Center, 2008), an open source software framework for the easy integration and utilization of datasets, algorithms, tools, and computing resources" [576].

NWB provides researchers and educators access to multiple algorithms and datasets. NWB allows users to make successive algorithm calls to create what is called a Workflow. The Workflow is designed to handle the application end-to-end, providing a framework for Data collection, Preprocessing, Analysis, Modelling and Visualization. All these operations are done with ease via the user-friendly Console window. The code library of NWB is basically just insertion of plug-ins. Some plug-ins run the core architecture, while additional plugins are also available for other problem specific algorithms. All the derived and the parent datasets are displayed in the Data Manager window. NWBs dexterity in data manipulation is exemplified by the Data Conversion Service which facilitates the conversion between different file types. Models such as Random Graph Model, Watts-Strogatz Small World etc. can be generated using NWB, thus facilitating the creation of descriptive models of compatible datasets. Tree Visualizations such as Tree View Visualization, Tree Map Visualization, Balloon Graph Visualization etc.

can be constructed over datasets which are natively in the tree format. The GUI is user-friendly, allowing users to set parameters such as size, color and shape for both the nodes and the edges; meanwhile, the legend is auto-generated. NWB also has the capability to produce various Graph Visualizations. Some of the Graph Visualizations available are LaNet, JUNG-based Circular. Other than the built-in ability to create visualizations, NWB also provides some integration with other tools such as GUESS and Gnuplot. The whole network visualization or any specific view can be saved in formats such as pdf, gif, raw, jpg or png [576].



title	OCCI
status	10
section	Interoperability
keywords	Interoperability

OCCI stands for Open Cloud Computing Interface [[ref missing](#)]. OCCI is a RESTful protocol and an Application Programming Interface (API) for all kinds of management tasks [577]. Originally OCCI was created for IaaS model-based services. The focus of the API was to allow for interoperability between tools that were aiming to accomplish similar tasks. Since its creation, OCCI has been transformed into a more flexible and broad reaching API that allows for future proofing but continues its tradition of aiming for interoperability. The OCCI specification also aims for portability and ease of integration in future systems. OCCI's currently release is useful to serve many different models including IaaS, PaaS, and SaaS.

OCCI's shift to be more modular and extensible requires the specification to be released in a suite of complimentary documents [577]. The current version of OCCI's specifications is 1.2. The documents that are included in the OCCI specification are split into 8 different documents, they are as follows: Core, Compute Resource Templates Profile, HTTP Protocol, Infrastructure, JSON Rendering, Platform, Service Level Agreements, and Text Rendering. Previously the specification was only split into three different documents, but as this specification has been shifting to be more extensible and modular the number of documents has increased.

The Core document goes into detail about the core model of OCCI. The Compute Resource Templates Profile document discusses best practices and how to achieve interoperability while utilizing OCCI. The HTTP document provides documentation on how to interface with

the OCCI model by using the OCCI API. OCCI's Infrastructure document gives a detailed explanation of the infrastructure for the IaaS domain which OCCI was originally designed for. The JSON Rendering document provides details regarding how to utilize this protocol with JSON Rendering. The Platform document provides details about the extension for the PaaS domain. Service Level Agreements documentation discusses how OCCI handles service level agreements. Finally, the Text Rendering document discusses how to interact with the OCCI protocol using TEXT rendering [578].

3.262 ODBC/JDBC



title	ODBC/JDBC
status	90
section	Object-relational mapping
keywords	Object-relational mapping

Open Database Connectivity (ODBC) is an open standard application programming interface (API) for accessing database management systems (DBMS) [579]. ODBC was developed by the SQL Access Group and released in September, 1992. Microsoft Windows was the first to provide an ODBC product. Later the versions for UNIX, OS/2, and Macintosh platforms were developed. ODBC is independent of the programming language, database system and platform.

Java Database Connectivity (JDBC) is a API developed specific to the Java programming language. JDBC was released as part of Java Development Kit (JDK) 1.1 on February 19, 1997 by Sun Microsystems [580]. The 'java.sql' and 'javax.sql' packages contain the JDBC classes. JDBC is more suitable for object oriented databases. JDBC can be used for ODBC compliant databases by using a JDBC-to-ODBC bridge.

3.263 ODE



title	ODE
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Apache ODE (Orchestration Director Engine) is an open source implementation of the WS-BPEL 2.0 standard. WS-BPEL which stands for Web Services Business Process Execution Language, is an executable language for writing business processes with web services [581]. It includes control structures like conditions or loops as well as elements to invoke web services and receive messages from services. ODE uses WSDL (Web Services Description Language) for interfacing with web services [582]. Naming a few of its features, It supports two communication layers for interacting with the outside world, one based on Axis2 (Web Services http transport) and another one based on the JBI standard. It also supports both long and short living process executions for orchestrating services for applications [583].

3.264 Omid



title	Omid
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Omid is a

“flexible, reliable, high performant and scalable ACID transactional framework” [584]

for NoSQL databases, developed by Yahoo for HBase and contributed to the Apache community. Most NoSQL databases, do not natively support ACID transactions. Omid employs a lock free approach from concurrency and can scale beyond 100,000 transactions per second. At Yahoo, millions of transactions per day are processed by Omid [585].

Omid is currently in the Apache Incubator. All projects accepted by the Apache Software Foundation (ASF) undergo an incubation period until a review indicates that the project meets the standards of other ASF projects [584]

3.265 OODT



title	OODT
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

The Apache Object Oriented Data Technology (OODT) is an open source data management system framework. OODT was originally developed at NASA Jet Propulsion Laboratory to support capturing, processing and sharing of data for NASA's scientific archives. OODT focuses on two canonical use cases: Big Data processing and on Information integration. It facilitates the integration of highly distributed and heterogeneous data intensive systems enabling the integration of different, distributed software systems, metadata and data. OODT is written in the Java, and through its REST API used in other languages including Python [586].

3.266 Oozie



title	Oozie
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Oozie is a workflow manager and scheduler. Oozie is designed to scale in a Hadoop cluster. Each job will be launched from a different datanode [587] [588]. Oozie is architected from the ground up for large-scale Hadoop workflow [589]. Scales to meet the demand, provides a multi-tenant service, is secure to protect data and processing, and can be operated cost effectively. As demand for workflow and the sophistication of applications increase, it must continue to mature in these areas [587]. Is well integrated with Hadoop security. Is the only workflow manager with built-in Hadoop actions, making workflow development, maintenance and troubleshooting easier. It's UI makes it easier to drill down to specific errors in the data nodes. Proven to scale in some of the world's largest clusters [587]. Gets callbacks from MapReduce jobs so it knows when they finish and whether they hang without expensive polling. Oozie Coordinator allows triggering actions when files arrive at HDFS. Also supported by Hadoop vendors [587].

3.267 Open MPI



title	Open MPI
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

The Open MPI Project [590] is an open source Message Passing Interface implementation that is developed and maintained by a consortium of academic, research, and industry partners. Open MPI is therefore able to combine the expertise, technologies, and resources from all across the High Performance Computing community in order to build the best MPI library available. Open MPI offers advantages for system and software vendors, application developers and computer science researchers. Open MPI [591] provides functionality that has not previously been available in any single, production-quality MPI implementation, including support for all of MPI-2, multiple concurrent user threads, and multiple options for handling process and network failures.



title	OpenCV
status	90
section	Application and Analytics
keywords	Application and Analytics

OpenCV [592] stands for Open Source Computer Vision which library most commonly used for detecting anything related to images. With more than 2500 algorithms inside the library, OpenCV is a type of machine learning framework that works with image files such as face or object recognition, moving videos where human or objects are constantly shifting each frame, 3D objects both for modelling and constructing, image stitching, and other possible frameworks working with images. As such, OpenCV is available on multiple platforms - supporting popular operating systems such as Windows, Linux, Android, and Mac -and API for languages such as C++, Python and Java [593].

Having a strong image analysis library, some of the use cases for OpenCV are Image Enhancement, Wiener Deconvolution, Hough Circles, Mean Shift algorithm, Image Segmentation, Image Blending, etc. For instances, image enhancement utilizing OpenCV allows users to improve the image quality from original source. Image is fed into the framework and will be enhanced through equalization of the RGB value. Another example would be mean shift algorithm which commonly used for tracking movement. The framework first recognized the object which the user will direct on the image, and store the attributes of the object which will then analyzed as histogram. The framework then follows the movement of the object and returns the coordinates of the moving object [594].

Besides image analysis, OpenCV also provides statistical machine learning libraries which become the backbone of the library itself.

Those libraries include Boosting, Decision tree learning, Gradient boosting, k-nearest algorithm, Naive Bayes classifier, artificial neural networks, support vector machine, and deep neural networks [595].



title	OPeNDAP
status	100
section	File management
keywords	File management

OPeNDAP is an acronym for

“Open-source Project for a Network Data Access Protocol” [596].

The OPeNDAP focused on enhancing the retrieval of remote, structured data through a Web-based architecture and a discipline-neutral Data Access Protocol (DAP) [596]. It allows users to access data anywhere from the internet using a variety of client/server methods, including Ferret. Employing technology similar to that used by the World Wide Web, OPeNDAP and Ferret create a powerful tool for the retrieval, sampling, analyzing and displaying of datasets; regardless of size or data format (though there are data format limitations) [597]. “OPeNDAP” often is used in place of “DAP” to denote the protocol but also may refer to an entire DAP-based data-retrieval architecture [596].

OPeNDAP is a framework that simplifies all aspects of scientific data networking. OPeNDAP provides software which makes local data accessible to remote locations regardless of local storage format. OPeNDAP also provides tools for transforming existing applications into OPeNDAP clients (i.e., enabling them to remotely access OPeNDAP served data). OPeNDAP software is freely available [598]. It widely used, especially in Earth science, the protocol is layered on HTTP, and its current specification is DAP4, though the previous DAP2 version remains broadly used. Developed and advanced (openly and collaboratively) by the non-profit OPeNDAP, Inc., DAP is intended to

enable remote, selective data-retrieval as an easily invoked Web service. OPeNDAP, Inc. also develops and maintains zero-cost (reference) implementations of the DAP protocol in both server-side and client-side software [596].

OPeNDAP provide data services based on cloud computing technology that is equivalent to those developed for traditional computing and storage systems is critical for successful migration to cloud-based architectures for data production, scientific analysis and storage [598].

OPeNDAP Web-service capabilities (comprising the Data Access Protocol (DAP) specification plus open-source software for realizing DAP in servers and clients) are among the most widely deployed means for achieving data-as-service functionality in the Earth sciences. OPeNDAP services are especially common in traditional data center environments where servers offer access to datasets stored in (very large) file systems, and a preponderance of the source data for these services is being stored in the Hierarchical Data Format Version 5 (HDF5) [599].

DODS/OPeNDAP driver implements read-only support for reading feature data from OPeNDAP (DODS) servers. It is optionally included in OGR if built with OPeNDAP support libraries. When opening a database, its name should be specified in the form “DODS:url”. The URL may include a constraint expression a shown here. Note that it may be necessary to quote or otherwise protect DODS URLs on the command line if they include question mark or ampersand characters as these often have special meaning to command shells. By default top level Sequence, Grid and Array objects will be translated into corresponding layers. Sequences are (by default) treated as point layers with the point geometries picked up from lat and lon variables if available. To provide more sophisticated translation of sequence, grid or array items into features it is necessary to provide additional information to OGR as DAS (dataset auxiliary information) either from the remote server, or locally via the AIS mechanism [600].



title	OpenID
status	10
section	Monitoring
keywords	Monitoring

3.270.1 Old Text

OpenID is an authentication protocol that allows users to log in to different websites, which are not related, using the same login credentials for each, i.e. without having to create separate id and password for all the websites. The login credentials used are of the existing account. The password is known only to the identity provider and nobody else which relieves the users' concern about identity being known to an insecure website [601]. It provides a mechanism that makes the users control the information that can be shared among multiple websites. OpenID is being adopted all over the web. Most of the leading organizations including Microsoft, Facebook, Google, etc. are accepting the OpenIDs [602]. It is an open source and not owned by anyone. Anyone can use OpenID or be an OpenID provider and there is no need for an individual to be approved.

3.270.2 New Text

OpenID [603] is an authentication protocol service that is intended to ensure identity protection for usage in software and corporate environments. A goal of OpenID is to fight both sides of the battle between convenience and security when it comes to information and its protection and privacy. Software such as OpenID is essential to ensure the integrity of data and protect against any attacks that may come from unwanted sources.

The technology uses end-user interactions with a relying party with

the intention of validating that user's identity. This service uses the internet and DNS services to identify source and destination of information and interactions with datasets, among other applications of the software. It utilizes several checksum verification processes and identifiers to ensure a user's identity.

There have been many attempted attacks on OpenID and it has been found to be susceptible to many kinds of phishing attacks [604]. Successful highjacking efforts have taken place over unsecured connections [604].

Data is stored in a JSON format which further enables usability by JavaScript clients and users [603]. Encryption, identity protection, and session management are also supported. There is a management console that alerts administrators of faults or security breaches as they are recognized.

OpenID came to respect in 2005 after it initially being named Yadis (Yet another distributed identity system) [603]. It has powerful connections and similarities to other identity authorities to ensure that the newest and most secure technologies are used and implemented correctly. OpenID's biggest competitor is a service called OAuth, which is less intensive on user information and does not qualify as an authentication protocol. OpenID supports many applications, including REST-like queries.



title	OpenJPA
status	10
section	Object-relational mapping
keywords	Object-relational mapping

3.271.1 NEW TEXT

Java Persistence Application [605] is a collection of devices and methods used to manage different data bases connected to Java [606]. JPA is a specification that needs implementation by using other applications. The use of JPA enables an organization to store and manage large amounts of information. The ability of storing large data bases of information is due the ability of the application's configurations to be used automatically in a large storage area.

Being a part of the larger spring data family, the Java Persistent makes it easier to access other data applications and data technologies. The Java application is important in compacting large amounts of information into a sizable one. Similarly, the use of Java persistence can also ensure easy access to information ensuring improved data access by reducing or flattening the processes required.

The use of the Java persistence application can only be done through the use of subscribers such as the Eclipse link. After identifying the suitable link, a JPA project has to be created. This project contains all the Java data bases and other devices connected to Java. Thereafter, one can create a new Java class and subsequently add it to the persistence file.

The use of the JPA among users has created a satisfactory effect among users. Most users find it a reliable option as opposed to the writing down of the different databases. Its effectiveness is due to its

ability to use mappings instead of writings, which have reduced so many writings of the data base. To conclude, most user's experience is positive towards the use of JAP makes it easy to load and save objects without necessarily using a specific language. Thus, Java persistence has created satisfaction among its consumers.



o: grammar

title	OpenNebula
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

for consistency use italic do thi in your other contributions also

OpenNebula [ref missing] is an open source cloud OS platform mainly focuses on data center virtualization to build infrastructure as a service (IaaS) for private, public and hybrid clouds [607] [608].

OpenNebula use monitoring, virtualization, storage, image, network, and authentication drivers to interact with the underlying technologies. OpenNebula's core parts described for list here all the keywords.

The VM manager is responsible for creating, deploying, suspending and shutting down the VM. It uses Xen, KVM and VMware hypervisor drivers to perform these operations. It also has mechanisms to ensure high availability to detect crashes and auto-restarting in case of failure occurs.

The network manager is responsible for creating private networks to connect internal components and managing public IP pools to make front-end services accessible to the outside. It needs network drivers to manage virtual networks on the physical network. It provides automatic MAC and IP address assignment and guarantees the traffic isolation between virtual networks.

The storage manager provides highly available and reliable storage service. It hides underlying physical storage details from the user and

makes the storage system easy manageable.

The image manager is responsible for creating, deleting, cloning VM images, and listing the current images. Users can share the images which they created to other users or publish it for public use.

The information manager monitors and provides information about the system including VM states, physical servers, and underlying devices. VM monitoring depends on the used hypervisor, so each hypervisor might not provide the same information.

"The federation manager enables access to remote cloud infrastructures, which can be either partner infrastructures governed by a similar cloud OS entity or public cloud providers. The federation manager should provide basic mechanisms for deployment, runtime management, and termination of virtual resources in remote clouds; remote resource monitoring; user authentication in remote cloud instances; access control management and remote resource permission; and tools for image building on different clouds with different image formats." [fa18-523-68-Moreno_Vozmediano-IaaS_Cloud:2012]

The service manager is responsible for deploying, suspending, resuming, and canceling multitier services. Multitier services have interconnected VMs and specific dependencies. Service manager checks service requirements and decides to accept or reject.

Apart from the managers, OpenNebula provides an authentication mechanism to manage access and accounting system to provide resource usage information [fa18-523-68-Moreno_Vozmediano-IaaS_Cloud:2012].

3.273 OpenPBS



title	OpenPBS
status	90
section	Cluster Resource Management
keywords	Cluster Resource Management

Portable Batch System (or simply PBS) is the name of computer software that performs job scheduling. Its primary task is to allocate computational tasks, i.e., batch jobs, among the available computing resources. It is often used in conjunction with UNIX cluster environments [609]. OpenPBS is the original open source version of PBS. There are more commercialized versions of the same software. One of the key features of OpenPBS is that it supports millions of cores with fast job dispatch and minimal latency. It meets unique site goals and SLAs by balancing job turnaround time and utilization with optimal job placement. OpenPBS also includes automatic fail-over architecture with no single point of failure - jobs are never lost, and jobs continue to run despite failures. It is built upon a Flexible Plugin Framework which simplifies administration with enhanced visibility and extensibility [610].

3.274 OpenRefine



title	OpenRefine
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

OpenRefine (formerly GoogleRefine) is an open source tool that is dedicated to cleaning messy data. With the help of this user-friendly tool you can explore huge data sets easily and quickly even if the data is a little unstructured. It allows you to load data, understand it, clean it up, reconcile it, and augment it with data coming from the web [611]. It operates on rows of data which have cells under columns, which is very similar to relational database tables. One OpenRefine project is one table. The user can filter the rows to display using facets that define filtering criteria. most operations in OpenRefine are done on all visible rows: transformation of all cells in all rows under one column, creation of a new column based on existing column data, etc. All actions that were done on a dataset are stored in a project and can be replayed on another dataset. It has a huge community with lots of contributors meaning that the software is constantly getting better and better.



title	OpenStack Heat
status	10
section	DevOps
keywords	DevOps

Openstack Heat [612], is a cloud deployment service was created as the main project by Openstack Orchestration Program with the mission of easily managing the various stages in the life-cycle of an application in a cloud environment and the infrastructure required in these different stages. One of the most important components in Cloud Computing environments is how the resources are managed, allocated. This process is known as Orchestration. This Heat project is essentially an orchestration service which allows us to define resources over the cloud and connections amongst them using a simple text file called referred as a template. This Heat template describes the infrastructure for a cloud application in a text file format that is easily understood by humans like code and allows changes to this template. Version Control is built into the Heat project to track any and all changes made to this Heat template . This template dynamically manages resources needed for our application. After the environment setup has been completed and is ready for execution, if there is a need to add or remove resources in an existing infrastructure this can be done using the template. The Heat program shall make all the necessary changes if and when the template being used is modified and the changes are pushed to main repository using the version control.

The heat-engine is the main component of the Openstack Heat architecture [613], it handles the bulk of the operations as a part of the orchestration process using the Heat template provided earlier by the developer . This heat-engine also communicates with the developer using the API (Application Program Interface) events. These

are the other useful components [614] in the Heat architecture, they are:

- Resources: These are execution objects that are created during runtime when the Orchestration process has started by the Heat program. Most of the infrastructure used in a cloud application could be considered a resource
- Stack: A Stack is the collection of various resources
- Parameters: These are fields that are used for dynamic input from the user during runtime or execution
- Output: These are fields that show output to the user

3.276 OpenStack Ironic



title	OpenStack Ironic
status	10
section	DevOps
keywords	DevOps

Ironic project is developed and supported by OpenStack. Ironic provisions bare metal machines instead of virtual machines and functions as hypervisor API that is developed using open source technologies like Preboot Execution Environment (PXE), Dynamic Host Configuration Protocol (DHCP), Network Bootstrap Program (NBP), Trivial File Transfer Protocol (TFTP) and Intelligent Platform Management Interface (IPMI) [615]. A properly configured Bare Metal service with the Compute and Network services, could provision both virtual and physical machines through the Compute service's API. But, the number of instance actions are limited, due to physical servers and switch hardware. For example, live migration is not possible on a bare metal instance. The Ironic service has five key components. A RESTful API service, through which other components would interact with the bare metal servers, a Conductor service, various drivers, messaging queue and a database. Ironic could be integrated with other OpenStack projects like Identity (keystone), Compute (nova), Network (neutron), Image (glance) and Object (swift) services.

title	OpenStack Keystone
status	10
section	Monitoring
keywords	Monitoring

“Keystone, the OpenStack Identity Service” [616]

OpenStack is a cloud service which falls under Infrastructure as a Service category (IaaS). It provides a combination of software tools for building infrastructure on clouds. OpenStack keystone is one of the main components of OpenStack architecture. The main purpose is to provide a high level authorization and authentication not only to users but also to OpenStack services [617]. Keystone authenticates users to avail other services such as image, computing, network, storage or dashboard from Openstack architecture by asking for credentials. It implements OpenStack’s Identity API and provides API client authentication, service discovery & distributed multi-tenant authorization.

Being an open source software, the source code is accessible to anyone who opts to use it. One of the most important advantages of OpenStack is that it allows users to deploy virtual machine along with other instances handling various tasks in managing a cloud environment. With an increase in the number of instances deployed by OpenStack, you can easily serve more number of users through tasks that can handle concurrent users.

One can add additional components to OpenStack and can thus customize it to their needs. Keystone is nothing but a project for OpenStack Identity that works with token, catalog, policy and assignment services through an OpenStack Application Programming Interface (API). Token validates and manages user token for

authentication purpose. Catalog provides end points registry which is used for endpoints discovery. Every service in OpenStack is connected through end points. Catalog gives you a general overview of connection of the users to the services. Policy provides rule base authorization which is associated with rule management. It checks for any kind of legal violations. Assignment services provide data about role. It authorizes different users based on their level of authorization [618]. New AUTH mechanisms such as oAuth, SAML and openID are included in the future versions of the Keystone component along with some proxying external services.

3.278 OpenStack



title	OpenStack
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

OpenStack [619] is a free and open source cloud operating system mostly deployed as infrastructure as a service (IaaS) that allows us to control large pool of computers, storage, and networking resources. OpenStack is managed by OpenStack Foundation [620].

Just like cloud, OpenStack provides infrastructure which runs as platform upon which end users can create applications. Key components of OpenStack include: Nova: which is the primary computing engine, Swift: which is a storage system for object and files, Neutron: which ensures effective communication between each of the components of the OpenStack. Other components include: Cinder, Horizon, Keystone, Glance, Ceilometer and Heat. The main goal of Openstack is to allow business to build Amazon-like cloud services in their own data centers. OpenStack is licensed under the Apache 2.0 license [621].

3.279 OpenTOSCA



title	OpenTOSCA
status	90
section	DevOps
keywords	DevOps

The Topology and Orchestration Specification for Cloud Applications, TOSCA is a new standard facilitating platform independent description of Cloud applications. OpenTOSCA is a runtime for TOSCA-based Cloud applications. The runtime enables fully automated plan-based deployment and management of applications defined in the OASIS TOSCA packaging format CSAR, Cloud Service ARchive. The key tasks of OpenTOSCA, are to operate management operations, run plans, and manage state of the TOSCA [622].



title	OpenVZ
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

OpenVZ is an open source operating system-level virtualization technology for the Linux kernel [623]. Analogous to many OS level virtualization technologies, OpenVZ lets multiple isolated Linux containers or Virtual Private Servers (VPSs) run on a single server. The VZkernel is a modified kernel which offers the host's functionality to each of the running isolated containers as the kernels don't have a dedicated kernel in place. When compared to hypervisor-based virtualization, OS based virtualization provides a smaller virtualization layer between containers and host thereby making the containers more elastic[624]. Only 1-2% of CPU resources are used on virtualization[625]. In openVZ the containers can be initiated by choosing OS templates that are available on OpenVZ website and each one of them is allotted a CID or container ID which uniquely define them and makes managing them easier. Also, multiple containers can be created by downloading single template. Each of the containers once executed, act like stand-alone servers. The templates selected to initiate containers, should have the same architecture as the host to run on the kernel. When multiple containers are executed in parallel, they share the same kernel although they have separate IP addresses, users, devices, file system etc... The allotment of resources like CPU, RAM etc. are taken care of resource management subsystem which has three important components :

1. Two-level disk quota : each container has its own disk quota set and internal container quota is set by its administrator
2. Fair CPU scheduler : Similarly to the disk quota allocation,

CPU scheduler has a two level allocation process. The containers are given a time sliced based on their CPU priority and internally the linux scheduler prioritizes the processes.

3. User BeanCounters : Makes sure that no container exploits the resources by having multiple counters, limits and guarantees in place to keep a check on each of the containers[626].

Also, an isolated container can be migrated from one server to another without any disruption in the containers work, this feature is called live migration which was added in 2006. This was achieved by the process called checkpointing, i.e., by freezing the container and saving it in a file on the disk, which then is sent to a different server where its restored. The whole process only happens in a matter of seconds. With such features, OpenVZ is used by institutions that can provide every user with a personal server and IT companies can use it for testing purposes.Infact,

> “OpenVZ can be efficiently applied in a wide range of areas : web hosting, enterprise server consolidation, software development and testing, user training, and so on” [627].

3.281 Oracle PGX



title	Oracle PGX
status	90
section	Application and Analytics
keywords	Application and Analytics

Numerous information is revealed from graphs. Information like direct and indirect relations or patterns in the elements of the data, can be easily seen through graphs. The analysis of graphs can unveil significant insights. Oracle PGX (Parallel Graph AnalytiX) is a toolkit for graph analysis.

“It is a fast, parallel, in-memory graph analytic framework that allows users to load up their graph data, run analytic algorithms on them, and to browse or store the result” [628].

Graphs can be loaded from various sources like SQL and NoSQL databases, Apache Spark and Hadoop [629].

3.282 Oracle fa18-423-06



title	Oracle
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Oracle Database ??? is a “relational database management system from the Oracle Corporation.” The software was originally developed by famous billionaire and founder of Oracle, Lawrence Ellington and other developers in 1977 ???. Oracle DB is a very popular database software that is very scalable and can be used across a wide variety of networks and that is one of the main reasons that Oracle DB is used by many major companies across the globe.

The power and usability of Oracle DB was seen in the Prime Day fiasco that hit Amazon in 2018. Amazon, who is developing their own personal database software, wants to be out of the Oracle DB system before 2020 due to arguments about pricing ???. However, their own software left them with slowed sales and thousands of delayed packages because of website and warehouse database errors. Matt Caesar, a computer science professor at the University of Illinois at Urbana-Champaign, believes “it appears they would have been able to diagnose the problem sooner if they were using Oracle’s database, which could possibly have reduced the outage duration” ???. Oracle DB has been used as a reliable and important software for years. The Amazon Prime Day mistake proves that the software is not only important, but necessary to the core functions of most large corporations.

You would think that the databasemarket is something that is very much “set-in-stone,” but Oracle has recently announced that they are rolling out an autonomous, cloud database that is gaining traction. An autonomous database can continuously tune and upgrade data

without downtime ????. The software is performing tasks autonomously that Database Managers would normally perform, such as updating the software, scaling-up data, and applying patches. One example is security patches. Security patches protect the data that is stored in the database from outside threats. According to Forbes, the number one cause of data breaches in companies is a system that lacks up-to-date securities patches ????. Therefore, the autonomous database not only automates redundant tasks such as upgrades and scalability, but protects the data from outside threats, an incredibly important issue in today's world.

3.283 ORC



title	ORC
status	90
section	File management
keywords	File management

ORC files were created as part of the initiative to massively speed up Apache Hive and improve the storage efficiency of data stored in Apache Hadoop. ORC is a self-describing type-aware columnar file format designed for Hadoop workloads. It is optimized for large streaming reads, but with integrated support for finding required rows quickly. Storing data in a columnar format lets the reader read, decompress, and process only the values that are required for the current query. Because ORC files are type-aware, the writer chooses the most appropriate encoding for the type and builds an internal index as the file is written. ORC files are divided into stripes that are roughly 64MB by default. The stripes in a file are independent of each other and form the natural unit of distributed work. Within each stripe, the columns are separated from each other so the reader can read just the columns that are required [630].



o: try to start with a sentence you wrote
about the tech not a quote

title	OSGi
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

OSGi [631] is a standard which focuses on providing a universal gateway for the users. It has been widely used as a framework to integrate all the services for a home. OSGi is widely used by business unit in many specific domains. Starting out in set-top boxes and residential gateways OSGi has today been adopted for solutions in IoT, M2M, Smart Home, Telematics, Assisted Living, Healthcare, Automotive, Media, Control Systems, Energy Management, Smart Meters, Telecommunications, Enterprise Software Platforms, and Robotics, to name a few.

"The OSGi Alliance, formerly known as the Open Services Gateway initiative, is a open standards organization founded in March 1999 that originally specified and continues to maintain the OSGi standard" [631].

In 1999, the original name of the organization was Open Service Gateway initiative, the aim was to offer service for the domestic gateway market. As more and more applications of Java have emerged, they decided to change the name to OSGi instead of Open Service Gateway initiative, and the original name should no longer be used.

The essence of OSGi is a modular system and a service platform for Java, which could offer a useful model. The functions do not existed in

the current system. The function is about teh operation the applications remotely and could do everything with the applications such as start, stop, update or something else. The APIs could also manage the application lifecycle through the protocol which designed for the application.

if its nowadays there must be a different refernce ...

The advantages of OSGi is that it could make the development become simpler, which could be a good news for development's cost budget, the whole amount of the project could be depressed to a lower level than before. For developers, the benefit is obvious that bugs are easy to find, the code is less to write, the structure could be more easy to design, the result could be more clear about the running process.

A series of specifications constitute OSGi, a reference implementation for each specification. For the programming language, OSGi could offer a dynamic module system which is made up by some specifications. As versions updating, OSGi is a mature platform to develop complex projects and products in many emerging domains.

3.285 Parasol



title	Parasol
status	90
section	Application and Analytics
keywords	Application and Analytics

The parasol laboratory is a multidisciplinary research program founded at Texas A&M University with a focus on next generation computing languages. The core focus is centered around algorithm and application development to find solutions to data concentrated problems [632]. The developed applications are being applied in the following areas: computational biology, geophysics, neuroscience, physics, robotics, virtual reality and computer aided drug design (CAD). The program has organized a number of workshops and conferences in the areas such as software, intelligent systems, and parallel architecture.

3.286 Parquet



title	Parquet
status	90
section	File management
keywords	File management

Apache parquet is the column Oriented data store for Apache Hadoop ecosystem and available in any data processing framework, data model or programming language [633]. It stores data such that the values in each column are physically stored in contiguous memory locations. As it has the columnar storage, it provides efficient data compression and encoding schemes which saves storage space as the queries that fetch specific column values need not read the entire row data and thus improving performance. It can be implemented using the Apache Thrift framework which increases its flexibility to work with a number of programming languages like C++, Java, Python, PHP, etc.



title	pbdR
status	10
section	Application and Analytics
keywords	Application and Analytics

Programming with Big Data in R (pbdR) is a series of R packages with S3/S4 objects and classes that are being used by the statisticians and data miners. pbdR deals with the data distributed on a series of machines in batch mode while R deals with a single machine. pdbr communicates between the machines using MPI's (Message Passing Interface).

There are two main implementations in R using MPI. They are Rmpi and pbdMPI of pbdR.

1. The pbdR built on pbdMPI uses SPMD (Single Program Multiple Data) parallelism where

“every processor is considered as worker and owns parts of data. There is no restriction to use manager/workers parallelism in SPMD parallelism environment” [634].

2. The Rmpi uses manager/workers parallelism where

“one main processor (manager) servers as the control of all other processors (workers)” [634].

pbdR not only works best for small data but also analyzing big data and more scalable for super computers that uses scalable linear algebra.

Programming with pbdR comprise of below packages:

pbdDEMO,pbdNCDF4,pbdDMAT,pmclust,pbdPROF,pbdZMQ,pbdMPI,p
pbdCS,kazaam,pbdRPC

"Among these packages, pbdMPI provides wrapper functions to MPI library, and it also produces a shared library and a configuration file for MPI environments. All other packages rely on this configuration for installation and library loading that avoids difficulty of library linking and compiling. All other packages can directly use MPI functions easily" [634].

R is an open source, and has a large user community. Users may extend the software by preparing contributed packages. Programming in R is through a simple and intuitive high level language, adapted from the S programming language, with rough similarity to Matlab programming. MPI (Message Passing Interface) is one of the most popular standards for general purpose distributed computing i.e, computing which is split and synchronized over multiple computers. MPI programs are traditionally written in lower level languages like C, C++, or FORTRAN. To write an MPI program, one should think from an SPMD (single program multiple data) perspective

3.288 Pegasus



title	Pegasus
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Pegasus is workflow management system that allows to compose and execute a workflow in an application in different environment without the need for any modifications [635]. It allows users to make high level workflow without thinking about the low level details. It locates the required input data and computational resources automatically. Pegasus also maintains information about tasks done and data produced. In case of errors Pegasus tries to recover by retrying the whole workflow and providing check pointing at workflow-level. It cleans up the storage as the workflow gets executed so that data-intensive workflows can have enough required space to execute on storage-constrained resources. Some of the other advantages of Pegasus are:scalability, reliability and high performance. Pegasus has been used in many scientific domains like astronomy, bioinformatics, earthquake science, ocean science, gravitational wave physics and others.

3.289 Pentaho



title	Pentah
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Pentaho is a business intelligence corporation that provides data mining, reporting, dashboarding and data integration capabilities. Generally, organizations tend to obtain meaningful relationships and useful information from the data present with them. Pentaho addresses the obstacles that obstruct them from doing so [636]. The platform includes a wide range of tools that analyze, explore, visualize and predict data easily which simplifies blending any data. The sole objective of pentaho is to translate data into value. Being an open and extensible source, pentaho provides big data tools to extract, prepare and blend any data [637]. Along with this, the visualizations and analytics will help in changing the path that the organizations follow to run their business. From spark and hadoop to noSQL, pentaho transforms big data into big insights.



title	PetSc
status	10
section	Application and Analytics
keywords	Application and Analytics

PetSc [638] stands for portable extensible toolkit for scientific applications. Scientific problems include lots of computations and require high level computational ability. PetSc provides a toolkit to solve such high complex problems, along with low level computing tools. Right now, it is available on many different platforms. In order to access those functions, one needs to include `petsc.h` in their program file (In C) and use those routines. It is also available in other languages such as C, C++, python and is supported on Linux, Unix, Mac Operating systems. Major components of PetSc includes NonLinear Equation Solvers, Time Stepping, Pre Conditioners, Krylov Subspace methods, Matrices, Vectors, IndexSets [639]. These components are utilized by application codes. PetSc also follows Object Oriented approach where there exists the objects such as Matrix, Linear Solver, NonLinear Solver. Developers coalesces with those objects of the classes that are required for their application.

PetSc is built on top of MPI, where it uses some of the capabilities of MPI (MPI provides the tools such as to perform operations on simple data types) where as PetSC provides us with intermediate tools. For instance, if one has to insert an element (matrix) in a random location or to perform matrix-vector product etc. It basically acts as an interface to softwares such as Plapack, Scalapack for Dense linear algebra, ParMetis, Party, Chaco, Jostle for Grid Partitioning softwares, IAO for Optimization, PVOODE for ODE solvers, SLEPc for Eigenvalue Solvers [640]. It provides tools for solving Linear system problems, distributed matrices and also extends support for profiling, debugging etc. It is widely used to solve these complex problems and

has wide community support. Moreover, this toolkit is free for anyone and there are many tutorials with examples and detailed documentation on the functions inside this library. The reason to build it in C rather than C++ or fortran is that, it has wide support and is very much evolved already and provides the capability to build data structures to deal with sparse matrices. Many ongoing scientific projects are utilizing Petsc libraries. It is also utilized to develop packages also used by algorithm developers and it greatly reduces the development time and effort.



title	Phoenix
status	100
section	High level Programming
keywords	High level Programming

In the first quarter of 2013, Salesforce.com released its proprietary SQL-like interface and query engine for HBase, Phoenix, to the open source community. The company appears to have been motivated to develop Phoenix as a way to (1) increase accessibility to HBase by using the industry-standard query language (SQL); (2) save users time by abstracting away the complexities of coding native HBase queries; and,(3) implementing query best practices by implementing them automatically via Phoenix [641]. Although Salesforce.com initially open-sourced it via Github, by May of 2014 it had become a top-level Apache project [642].

In the first quarter of 2013, Salesforce.com released its proprietary SQL-like interface and query engine for HBase, Phoenix, to the open source community. The company appears to have been motivated to develop Phoenix as a way to (1) increase accessibility to HBase by using the industry-standard query language (SQL); (2) save users time by abstracting away the complexities of coding native HBase queries; and, (3) implementing query best practices by implementing them automatically via Phoenix [641]. Although Salesforce.com initially open-sourced it via Github, by May of 2014 it had become a top-level Apache project [642].

Phoenix, written in Java,compiles [SQL queries] into a series of HBase scans, and orchestrates the running of those scans to produce regular JDBC result sets [643].

In addition, the program directs compute intense portions of the calls

to the server. For instance, if a user queried for the top ten records across numerous regions from an HBase database consisting of a billion records, the program would first select the top ten records for each region using server-side compute resources. After that, the client would be tasked with selecting the overall top ten [644].

Despite adding an abstraction layer, Phoenix can actually speed up queries because it optimizes the query during the translation process [641]. For example, 'Phoenix beats Hive for a simple query spanning 10M-100M rows' [645]. Another program can enhance HBase's accessibility for those inclined towards graphical interfaces. SQuirell only requires the user to set up the JDBC driver and specify the appropriate connection string [646].

The Apache Phoenix work as SQL skin for Hbase. Phoenix provides the flexibility to write queries like SQL when we are working on Hadoop API data. The Phoenix applications can run Map Reduce jobs as per user request and utilize the big data fundamentals. Apache Phoenix is increasing popularity over other tools available in its space. The beauty is that Phoenix provides features such as skipping full table scan, improve performance of overall system [647].

By utilizing HBase as its storage database, Phoenix enable OLTP and analysis for lower latency applications in Hadoop by combining standard SQL and JDBC APIs with full ACID transaction capabilities. The Phoenix support easy integration with other Hadoop ecosystem product like Hive [648], Pig [649], Map Reduce [647].

Phoenix framework provides the client and server libraries. Phoenix custom HBase co-processor handle metadata management, transaction, join, indexing, schema and on server side.

On Client end, Phoenix client library has parser, necessary relationship algebra and query plan component that used to parse the given query and choose the optimal plan based on cost-based optimization. Once query plan chooses, Phoenix internally convert the request to SCAN, PUT or DELETE operation and execute the operations [650].

The Java should be present on system with Hadoop to install Phoenix. The recent JDK V1.8.x JVM need for installation. The Hadoop and Phoenix can install on Windows, MAC, Linux systems [643].

The table creation and versioned control supported by Apache Phoenix Schema. The HBase table maintains table metadata and versioned. The snapshot queries will use the correct schema over prior versions.

A Phoenix table is created through the CREATE TABLE command and can either be:

1. built from scratch, in which case the HBase table and column families will be created automatically.
 2. mapped to an existing HBase table, by creating either a read-write TABLE or a read-only VIEW, with the caveat that the binary representation of the row key and key values must match that of the Phoenix data types.
- For a read-write TABLE, column families will be created automatically if they don't already exist. An empty key value will be added to the first column family of each existing row to minimize the size of the projection for queries.
 - For a read-only VIEW, all column families must already exist. The only change made to the HBase table will be the addition of the Phoenix co-processors used for query processing. The primary use case for a VIEW is to transfer existing data into a Phoenix table, since data modification are not allowed on a VIEW and query performance will likely be less than as with a TABLE [643].

3.292 Pig



title	Pig
status	10
section	High level Programming
keywords	High level Programming

3.293 Pilot Jobs



title	Pilot Jobs
status	90
section	Cluster Resource Management
keywords	Cluster Resource Management

In pilot job, an application acquires a resource so that it can be delegated some work directly by the application; instead of requiring some job scheduler. The issue of using a job scheduler is that a waiting queue is required. Few examples of Pilot Jobs are the [651] Falkon lightweight framework and [652] HTCaas. Pilot jobs are typically associated with both Parallel computing as well as Distributed computing. Their main aim is to reduce the dependency on queues and the associated multiple wait times.

Using pilot jobs enables us to have a multilevel technique for the execution of various workloads. This is so because the jobs are typically acquired by a placeholder job and they relayed to the workloads [653].

3.294 Pivotal Gemfire



title	Pivotal Gemfire
status	90
section	NoSQL
keywords	NoSQL

A real-time, consistent access to data-intensive applications is provided by a open source, data management platform named Pivotal Gemfire.

“GemFire pools memory, CPU, network resources, and optionally local disk across multiple processes to manage application objects and behavior”.

The main features of Gemfire are high scalability, continuous availability, shared nothing disk persistence, heterogeneous data sharing and parallelized application behavior on data stores to name a few. In Gemfire, clients can subscribe to receive notifications to execute their task based on a specific change in data. This is achieved through the continuous querying feature which enables event-driven architecture. The shared nothing architecture of Gemfire suggests that each node is self-sufficient and independent, which means that if the disk or caches in one node fail the remaining nodes remain untouched. Additionally, the support for multi-site configurations enable the user to scale horizontally between different distributed systems spread over a wide geographical network [654].

3.295 Pivotal GPLOAD/GPFDIST



title	Pivotal GPLOAD/GPFDIST
status	10
section	Data Transport
keywords	Data Transport

Greenplum Database is a shared nothing, massively parallel processing solution built to support next generation data warehousing and Big Data analytics processing [655]. In its new distribution under Pivotal, Greenplum Database is called Pivotal (Greenplum) Database.

gpfdist is Greenplum's parallel file distribution program [656]. It is used by readable external tables and gpload to serve external table files to all Greenplum Database segments in parallel. It is used by writable external tables to accept output streams from Greenplum Database segments in parallel and write them out to a file.

gpload is data loading utility is used to load data into Greenplum's external table in parallel [655].

Google has an invention relating to integrating map-reduce processing techniques into a distributed relational database. An embodiment of the invention is implemented by Greenplum as gpfdist [657].

3.296 Pivotal Greenplum



title	Pivotal Greenplum
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

Pivotal Greenplum is a commercial fully featured data warehouse. It is powered by Greenplum Database an open source initiative.

“It is powered by advanced cost-based query optimizer thereby delivering high analytical query performance on large data volumes”.

Pivotal Greenplum is uniquely focused on big data analytics [658].

The system consists of a master node, standy master node and segment nodes. The master node consists of the catalog information whereas the data resides on the segment nodes. The segment nodes runs on one or more segments which are modified PostgreSQL databases and are assigned a content identifier. The data is distributed among these segment nodes. The segment node also supports bult loading and unloading. The master node parses, optimizes an SQL query and dispatch it to all segment nodes. Therefore, it provides powerful and rapid analytics on petabyte scale data volumes [659].

3.297 Pivotal



title	Pivotal
status	90
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Pivotal Software, Inc. (Pivotal) is a software and services company. It offers multiple consulting and technology services, which includes Pivotal Web Services, which is an agile application hosting service. It has a single step upload feature cf push, another feature called Buildpacks lets us push applications written for any language like Java, Grails, Play, Spring, Node.js, Ruby on Rails, Sinatra or Go. Pivotal Web Services also allows developers to connect to 3rd party databases, email services, monitoring and more from the Marketplace. It also offers performance monitoring, active health monitoring, unified log streaming, web console built for team-based agile development [660].



title	PLASMA MAGMA
status	90
section	Application and Analytics
keywords	Application and Analytics

PLASMA stands for

“Parallel Linear Algebra Software for Multi-core Architectures” [661].

This technology is a successor of the two software libraries developed in the 80’s and 90’s named LAPACK and ScaLAPACK, respectively [661]. It represents one of the high-performance parallel programming models for dense linear algebra (DLA) that allows the

“solution of general systems of linear equations, symmetric positive definite systems, or linear equations and linear least squares problems, using LU, Cholesky, QR, and LQ factorizations” [662].

It was developed using tile algorithms and supports double and single precision. One of the basic concepts on which PLASMA was built is the unhidden parallelism contrary to the LAPACK parallelism which was hidden inside the Basic Linear Algebra Sub-programs (BLAS) [662]. One of the main reasons why PLASMA was developed was to help with LAPACK’s limitations when that technology is used on multicore processors. Even though PLASMA is more advanced than its predecessor, it still has a few limitations such as the inability to handle band matrices, eigenvalues, singular value problems, and computations on computers with distributed memory [662].

Another technology of this kind is MAGMA, which stands for

"Matrix Algebra on GPU and Multi-core Architectures" [661].

This library relies on LAPACK in functionality, data storage, and interface [663], however, contrary to its predecessor, it allows fast linear algebra computation on hybrid/heterogeneous architectures. Its hybrid algorithms

"rely on hybrid scheduler (of DAGs), hybrid kernels for nested parallelism, and existing software structure" [663].

This specific concept leverages the strengths of each individual hybrid component and represents the group of linear algebra algorithms as an assortment of tasks as well as the data dependencies among them [663]. In other words, it breaks down more demanding computations into multiple tasks of varying granularity and distributes their execution over multiple hardware components [663]. There are two different ways in which the execution of those tasks can be arranged - static and dynamic. More simple, smaller, and less demanding tasks are usually scheduled on CPU's, while larger, more demanding tasks are scheduled on the GPUs.

It is expected for these technologies to further advance with the development of the next generation of multi-core chips and accelerators (GPUs).

3.299 point-to-point



title	point-to-point
status	90
section	Inter process communication Collectives
keywords	Inter process communication Collectives



title	PolyBase
status	10
section	High level Programming
keywords	High level Programming

PolyBase is a technology which facilitates the direct query of an external distributed system such as Hadoop from SQL server using simple T-SQL queries. In most applications, data is almost always stored in multiple environments. With some of it being stored in standard relational databases and others being stored in Unstructured/semi-structured format in a Hadoop environment. For organizations to be able to analyze their data, conventionally, they would have to deal with complex MapReduce actions to be able to deal with the externally stored data. This can be quite inconvenient as this can only be done by a user with the knowledge of MapReduce. Before Polybase, in a situation where data is stored partially in a SQL server Database and partially in an external Hadoop environment, the way to perform calculations on the whole data would have been by transferring either part of the data into the other side so that they are in one single format. PolyBase provides an easy way to deal with the querying of the joined data without the need for additional ETL. The data from the External framework is not brought into the SQL server instance, instead, a standard T-SQL query is pushed into the Hadoop framework and only the result is returned. Hence PolyBase effectively eliminates the need for the user to have the knowledge of complicated MapReduce actions. By pushing the T-SQL into the Hadoop environment, the query load is transferred to the distributed system which tend to be much faster due to the distributed query processing, thus increasing the overall performance of the process. Another added advantage of PolyBase is that the entire process of querying is done without the need of any additional software on the Hadoop environment. Initially designed to be used only with SQL

Server Parallel Data Warehouse (PDW), PolyBase began being included with SQL Server from the 2016 edition. Apart from PDW and MS SQL Server, PolyBase is also compatible with Windows Azure Blob Store, which means that its functionality goes beyond dealing with just a HDFS on Hadoop system. PolyBase is also compatible with any BI tool compatible with SQL Server [664].

3.301 PostgreSQL



title	PostgreSQL
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

3.301.1 Old Entry

PostgreSQL is an open-source relational database management system (DBMS). It runs on all the major operating systems like Linux, Mac OSX, Windows and UNIX. It supports the ACID (Atomicity, Consistency, Isolation and Durability) properties of a conventional DBMS. It supports the standard SQL:2008 data types like INTEGER, NUMERIC, etc. besides providing native interfaces for languages such as C++, C, Java and .Net [665].

With the release of its latest version 9.5, it has included new features like the UPSERT capability, Row Level security and multiple features to support Big Data. These new features rolled out in the latest version make PostgreSQL a very strong contender for modern use. UPSERT feature has predominantly been released for the application developers in order to help them simplify their web application and software development. UPSERT is basically a shorthand of Insert, on conflict update. Row Level Security (RLS), as the name suggests, enables the database administrators to control which particular rows could be updated by the users. This helps in ensuring that the users do not inadvertently update rows which they are not meant to. Features such as BRIN indexing, Faster sorts, CUBE, ROLLUP and GROUPING SETS, Foreign Data Wrappers and TABLESAMPLE were added as a part of the new Big Data features. Under BRIN indexing (Block Range Indexing), PostgreSQL supports creating small but powerful indexes for large tables. Using a new algorithm called as abbreviated keys, PostgreSQL can sort NUMERIC data very quickly.

The CUBE, ROLLUP and GROUPING clauses enable the users to use just a single query to create myriad reports at different levels of summarization. Using the concept of Foreign Data Wrappers (FDWs), PostgreSQL can be used for querying Big Data systems like Cassandra and Hadoop. The TABLESAMPLE clause allows quick statistical sample generation of huge tables without any need to sort them [666].

3.301.2 New entry

PostgreSQL, often referred as Postgres [665], is an open source, object-relational database management system. PostgreSQL is free, extensible and supports cross platform feature. Its source code is available with open source licence. Postgres was created at UCB by a computer science professor named Michael Stonebraker [667].

PostgreSQL runs on all major operating systems. Initially it was designed to run on UNIX platforms. Now it works on 34 platforms of Linux along with other platforms such as all Windows versions, Mac OS X and Solaris. It supports text, images, sounds, video and includes programming interfaces for different languages such as C, C++, Java, Perl, Python, Ruby, Tcl and Open Database Connectivity.

PostgreSQL is completely ACID compliant and transactional. It has complete support for different features such as foreign keys, joins, views, triggers, and stored procedure [333]. It includes almost all data types that are used in SQL, such as INTEGER, NUMERIC, BOOLEAN, CHAR, VARCHAR, DATE, INTERVAL, and TIMESTAMP data type. It also supports storage of binary large objects, including pictures, sounds, or video [665].

3.302 Potree



title	Potree
status	10
section	Application and Analytics
keywords	Application and Analytics

Potree is a opensource tool powered by WebGL based viewer to visualize data from large point clouds [668]. It started at the TU Wien, institute of Computer Graphics and Algorithms and currently begin continued under the Harvest4D project. Potree relies on reorganizing the point cloud data into an multi-resolution octree data structure which is time consuming. Its efficiency can be improved by using techniques such as divide and conquer as discussed in a conference paper Taming the beast: Free and Open Source massive cloud point cloud web visualization [669]. It has also been widely used in works involving spatio-temporal data where the changes in geographical features are across time [670].



title	Pregel
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Pregel [671] is a system that is predominantly used for large-scale graph processing. It is a framework that is used to query distributed and directed graphs. It is also known as Map-Reduce for graphs because it uses the same phases i.e. it has a map phase and a reduce phase. It uses several iterations of map reduce i.e. several map reduce steps one after the other. It tries to operate all the servers that are present in a Hadoop cluster at full capacity at all the time whenever possible as well as it also reduces the network traffic which is the sole purpose of having a graph in a distributed cluster. It is particularly good at calculating measurements and values that are touching most of the vertices or all of the vertices in the graph and it is really bad at calculating values that touch just a few vertices .

The Pregel sequence has two primary components conductor and worker. The workers are located at the data on the database servers, but the conductor could be anywhere else i.e. it could be on the same server or a different server. The conductor acts as an interface between the user and the database server. The user simply inputs the required task to the conductor which transmits the same to the worker which compute over all the vertices of the graph and send messages between the vertices. The entire process is performed oblivious to the conductor.

The most important step in the map reduce process is the implementation of the worker algorithm and it is implemented by the

user. This implementation is delivered across all the database servers and is mapped over all the vertices of the graph. This part of the process is the map step since it maps the algorithm implementation over all the vertices and as an output it creates a set of messages to other services at other vertices. The reducer phase is composed of combiners which reduce all the generated messages from the worker and output an aggregated message at each vertex [672]. This process ensures aggregation of all the messages from one server to the other server.

3.304 Presto



title	Presto
status	10
section	High level Programming
keywords	High level Programming

Presto is an open-source distributed SQL query engine that supports interactive analytics on large datasets [673]. It allows interfacing with a variety of data sources such as Hive, Cassandra, RDBMSs and proprietary data source. Presto is used at a number of big-data companies such as Facebook, Airbnb and Dropbox. Presto's performance compares favorably to similar systems such as Hive and Stinger [674].



title	Protobuf
status	90
section	Message and Data Protocols
keywords	Message and Data Protocols, RPC, XML

Protocol Buffer is a way to serialize structured data into binary form in order to transfer it over wires or for storage [675]. It is used for inter application communication or for remote procedure call (RPC). It involves an interface description that describes the structure of some data and a program that can generate source code or parse it back to the binary form. It emphasizes simplicity, performance and speed. Protocol Buffer allows user to define the schema for their data and put messages within the schema [676]. Protocol Buffers has APIs for Python, Java, and C++ and other projects are in the works to implement other languages as well [677]. Protocol Buffer defines its schema definition in the .proto and is built with key/value pairs.

Protocol Buffers is compared frequently to

[678]. The tradeoff between XML and protocol buffers is read ability and speed/size. Protocol buffers are typically smaller, faster but require the consumer to have the protocol file to de-serialize the data [678]. XML provides the information needed to parse but is typically slower to parse and transmit because of its size difference to Protocol Buffers [679].

Protocol Buffers was developed by Google and by 2008 was primary in much of Google's infrastructure???. Google sought to make Protocol Buffers accessible, so in 2008 Google made Protocol Buffers available open source in hopes that invitestment from the open source community would make Protocol Buffers even better ???.

Protocol Buffers wasn't intended to deal with larger files [680]. Big data challenges frequently include larger files. However, Protocol

Buffers is ideal for smaller structured pieces of a larger message. For example, Google Maps files includes large amounts of data, but Protocol Buffers is the ideal solution because tranferring and loading a large amount of data in small pieces is what where Protocol Buffers can excel.

3.306 (b) publish-subscribe: Big Data



title	(b) publish-subscribe: Big Data
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

Publish/Subscribe (Pub/Sub) is a communication paradigm in which subscribers register their interest as a pattern of events or topics and then asynchronously receive events matching their interest [681]. On the other hand, publishers generate events that are delivered to subscribers with matching interests. In Pub/sub systems, publishers and subscribers need not know each other. Pub/sub technology is widely used for a loosely coupled interaction between disparate publishing data-sources and numerous subscribing data-sinks. The two most widely used pub/sub schemes are - Topic-Based Publish/Subscribe (TBPS) and Content-Based Publish/Subscribe (CBPS) [682].

Big Data analytics architecture are being built on top of a publish/subscribe service stratum, serving as the communication facility used to exchange data among the involved components [683]. Such a publish/subscribe service stratum brilliantly solves several interoperability issues due to the heterogeneity of the data to be handled in typical Big Data scenarios.

Pub/Sub systems are being widely deployed in Centralized datacenters, P2P environments, RSS feed notifications, financial data dissemination, business process management, Social interaction message notifications- Facebook, Twitter, Spotify, etc.



title	Puppet
status	10
section	DevOps
keywords	DevOps

Puppet [684] is an open-source configuration management system which is a server that holds all the configuration information for the different agents or the different servers that check into it . Devops is the recent trend where continuous integration and deployment has become more famous and is adopted by many companies. Configuration Management in terms of servers is when there are a lot of servers in a data center, an organization would want to keep these servers in a particular state. Puppet helps in automating the deployment of developed services and thus helps in developing the automated infrastructure [685]. This helps developers to concentrate more on the critical work other than the repetitive and monotonous work. Thus, it helps in being productive and spends less time in managing routine work. Puppet is a configuration management tool where there exists a puppet master and many puppet agents. One can store all the information regarding configuration in puppet master and puppet agent checks with it for the information [686]. For example, if one has to build 100's of systems, 20 with one configuration (one can specify to install different OS, RAM and softwares on top of it) and rest with another configuration. Imagine doing this manually, and individually building each server with the configuration specified. It is a very cumbersome process. Instead with the puppet, one can use a server where puppet master is installed and maintain different configuration files in it (This file specifies how each machine needs to be built). Now on other machines one can just install puppet agents. These puppet agents now pick up the configuration files from the puppet master server (each puppet agent picks corresponding configuration file that matches the mac address)

and each puppet agent installs/builds the systems based on the configuration file on puppet server [687]. Thus, one can achieve consistency across all the servers, make sure all of them are up to date.



title	PyBrain
status	10
section	Application and Analytics
keywords	Application and Analytics

Pybrain is an open source machine learning library for python [688]. Pybrain is a research project and it is developed by the researchers at the Dalle Molle Institute for Artificial intelligence in Switzerland and Technical university of Munich, Germany. Pybrain can be independently used and its only dependency is on SciPy.

“While there are a few machine learning libraries out there, PyBrain aims to be a very easy-to-use modular library that can be used by entry-level students but still offers the flexibility and algorithms for state-of-the-art research” [689].

PyBrain can compose custom neural network architectures that vary from recurrent networks that are multi-dimensional to Boltzmann machines also known as convolutional networks. And in this way, it is different from the other Machine Learning Libraries.

PyBrain can be extensively use for Supervised, Unsupervised, Semi supervised and Reinforcement learning and black box for parallel optimization. PyBrain also supports various Artificial neural networks. PyBrain has methods and sub library that supports plotting functions, read and write to the XMLs. We can do implicit mapping using PyBrain and use it as an implicit layer to build Gamming and 3D environments. It is also used to overcome the problems of continuous state and action spaces associated with the real-life tasks.

Pybrain includes various training algorithms, data handling tools and

environments that work with sequential and non-sequential data. PyBrain architectures includes Long Short-Term Memory (LSTM), Deep Belief Networks with the prominent ones being the Feedforward Network and Recurrent Network the Feedforward Network is nothing but an artificial neural network establishing a connection between the nodes unlike the Recurrent Networks where a set of vertices connected by the edges are formed along the sequence by connecting the nodes. It allows multiple algorithms to be incorporated together to get the best possible results. PyBrain stands out among other machine learning library in python because of its simplicity and convenient way of applying most advanced AI algorithm [688].

3.309 QEMU



title	QEMU
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

QEMU (Quick Emulator) is a generic open source hosted hypervisor [411] that performs hardware virtualization (virtualization of computers as complete hardware platform, certain logical abstraction of their componentry or only the certain functionality required to run various operating systems) [690] and also emulates CPUs through dynamic binary translations and provides a set of device models, enabling it to run a variety of unmodified guest operating systems.

When used as an emulator, QEMU can run Operating Systems and programs made for one machine (ARM board) on a different machine (e.g. a personal computer) and achieve good performance by using dynamic translations. When used as a virtualizer, QEMU achieves near native performance by executing the guest code directly on the host CPU. QEMU supports virtualization when executing under the Xen hypervisor or using KVM kernel module in Linux [691].

Compared to other virtualization programs like VMWare and VirtualBox, QEMU does not provide a GUI interface to manage virtual machines nor does it provide a way to create persistent virtual machine with saved settings. All parameters to run virtual machine have to be specified on a command line at every launch. It is worth noting that there are several GUI front-ends for QEMU like virt-manager and gnome-box.

3.310 QPid



title	QPid
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives



title	R
status	90
section	Application and Analytics
keywords	Application and Analytics

R [692] is a free implementation of the S programming language, which was originally created and distributed by Bell Labs. It can be used to perform anything from basic to advanced statistical and graphical techniques.

"R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis" [693].

The many advantages of R include its package ecosystem. Developers agree that if a statistical technique exists then there is a good chance that a package for this function already exists. Being open-source, r allows new packages to be written and shared among programmers. R can also be integrated with other programming languages like C and C++. There is however one disadvantage to using R - it can consume all available memory.

R code can be run without any compiler owing to the fact that it is an interpreted language. This makes the development of code easier. Being a vector language, you can add functions to a single Vector without implementing a loop. Hence, R is more faster than other languages. R is utilised in many fields from biology to genetics and statistics among others.

The question here would be why use R when software like MS Excel, Matlab, SAS etc exist. Its simple, they all have limitations varying from inability to handle very large datasets to being used for very specific scenarios and most importantly being expensive. R Tackes all these problems. R is good for exploring datasets and ad hoc analysis [694]. In addition to statistical analysis, R can also be used for Data Mining. The most common data mining techniques include pattern recognition, classification, clustering etc. These advantages over other statistical software encourage the growing use of R in cutting edge social science research.



o: grammar or factual inaccurate statements. At times the grammar makes it factual inaccurate, at other times the grammar is correct and the fact is questionable or unclear presented.

title RabbitMQ

status 10

section Inter process communication Collectives

keywords Inter process communication Collectives

RabbitMQ [695] is widely used for the management of big data. It is a message-queuing technology which is also referred to as a message manager. RabbitMq maintains a path from Producer to Consumer to ensure that data is transmitted in a specified way. Research indicates[696]that it is one of the most common message brokers. RabbitMQ works by receiving messages from producers and pushing them to queues based on the rules and policies that the exchange type outlines. It enables web servers to respond rapidly to requests instead of being forced to undertake resource heavy procedures abruptly.

It is lightweight and relatively easy to deploy on a devices as well as in a cloud. This technology is deployable in federated as well as distributed configurations to achieve high-availability and high-scale requests. Furthermore, it can run on most environments people are using today, is compatible with most operating systems, and offers assorted developer tools for most popular languages.



title	Rasdaman
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Rasdaman stands for raster data management. It is a database management system that facilitates the storage and retrieval of multi-dimensional arrays - like sensor, image etc[697]. It is a raster database middleware offering an SQL-style query language on multi-dimensional arrays of unlimited size, stored in a relational database. The first prototype was developed in TU Munich. Peter Baumann established a database model for multi-dimensional arrays, including a data model and declarative query language. Rasdaman is a Big Data Engine for flexible ad-hoc analytics on multi-dimensional spatio-temporal sensor, image, simulation, and statistics data of unlimited size. The Web Coverage Processing Service (WCPS) query language is an Open Geospatial Consortium (OGC) standard which allows filtering and processing of multi-dimensional raster coverages, such as sensor, simulation, image, and statistics data using web services. The WCPS queries are translated to rasdaman query language, rasql, and are executed on rasdaman. This quick start shows how to access and manipulate an example 2D coverage using WCPS language.

rasdaman supports open big data standards[698].

3.313.1 Key Features

1. **fast**: parallel access to Exascale archives and Terabyte objects in fractions of a second.
2. **scalable**: seamlessly from laptop to high-parallel, high-availability clouds and server farms.
3. **flexible**: Array SQL for navigation, extraction, processing, and

ad-hoc analysis. Array data can reside in a conventional database, in files optimized by rasdaman, or in some pre-existing archive.

4. **open standards as issued by OGC:** WMS, WCS, WCS-T, WCPS; rasdaman is WCS Core Reference Implementation and listed in the GEOSS Component and Service Registry.
5. **free:** available as open source in a lively, mature, and professionally managed open-source project supervised by Jacobs University, in incubation by the OSGeo foundation.
6. **cost-efficient:** through intelligent, economic resource utilization and free source code.

Raster objects are maintained in a standard relational database by partitioning of a raster object into tiles.

Any user or system partitioning can be generated. Tiles form the unit of disk access. The tiling pattern is adjusted to the query access patterns; several tiling strategies assist in establishing a well-performing tiling. A geo index is employed to quickly determine the tiles affected by a query. Sometimes the tiles are compressed using various algorithms such as lossless and lossy (wavelet) algorithms. Both tiling strategy and compression comprise database tuning parameters. Tiles and tile index are stored as BLOBs in a relational database which also holds the data dictionary needed by rasdaman's dynamic type system. Adapters are available for several relational systems, among them open-source PostgreSQL. For arrays larger than disk space, hierarchical storage management (HSM) support has been developed[699].

3.314 Razor



title	Razor
status	10
section	DevOps
keywords	DevOps

Razor is a hardware provisioning application, developed by Puppet Labs and EMC. Razor was introduced as open, pluggable, and programmable since most of the provisioning tools that existed were vendor-specific, monolithic, and closed. Razor can deploy both bare-metal and virtual systems. During boot the Razor client automatically discovers the inventory of the server hardware - CPUs, disk, memory, etc., feeds this to the Razor server in real-time and the latest state of every server is updated [700]. It maintains a set of rules to dynamically match the appropriate operating system images with server capabilities as expressed in metadata. User-created policy rules are referred to choose the preconfigured model to be applied to a new node. The node follows the model's directions, giving feedback to Razor as it completes various steps as specified in [701]. Models can include steps for handoff to a DevOps system or to any other system capable of controlling the node.

3.315 RCFile



title	RCFile
status	10
section	File management
keywords	File management

RCFile (Record Columnar File) is a big data placement data structure that supports fast data loading and query processing coupled with efficient storage space utilization and adaptive to dynamic workload environments [702]. It is designed for data warehousing systems that uses map-reduce. The data is stored as a flat file comprising of binary key/value pairs. The rows are partitioned first and then the columns are partitioned in each row and the respective meta-data for each row is stored in the key part for that row and the values comprises of the data part of the row. Storing the data in this format enables RCFile to accomplish fast loading and query processing. A shell utility is available for reading RCFile data and metadata [703]. RCFile has been chosen in Facebook data warehouse system as the default option [704]. It has also been adopted by Hive and Pig, the two most widely used data analysis systems developed in Facebook and Yahoo!

3.316 Red Hat OpenShift



title	Red Hat OpenShift
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

OpenShift was launched as a PaaS (Platform as a Service) by Red Hat in the Red Hat Summit, 2011 [705]. It is a cloud application development and hosting platform that envisages shifting of the developer's focus to development by automating the management and scaling of applications [706]. Thus, OpenShift enables us to write our applications in any one web development language (using any framework) and it itself takes up the task of running the application on the web [707]. This has its advantages and disadvantages - advantage being the developer doesn't have to worry about how the stuff works internally (as it is abstracted away) and the disadvantage being that he cannot control how it works, again because it is abstracted.

OpenShift is powered by Origin, which is in turn built using Docker container packaging and Kubernetes container cluster [708]. Due to this, OpenShift offers a lot of options, including online, on-premise and open source project options.



title	Redis
status	10
section	In-memory databases/caches
keywords	In-memory databases/caches

3.317.1 NEW TEXT

Redis [709] is an open source in-memory structured query language (SQL) that was developed by Salvatore Sanfilippo while working for VMware, which is a cloud infrastructure company [710]. Redis utilizes in-memory storage that makes a website processing speed to be faster. Moreover, its incorporation of Lua scripts adds flexibility to a user's experience.

Redis includes 5 data types, such as list, set, string, sorted set, and hash. Moreover, it has other two special data types namely hyper log-log and bitmap. Secondly, the system has 160 commands that one can implement to make transactions. Hence, these commands enable a user to have a variety of specialized custom commands. Thirdly, Redis is single-threaded, which means that it is capable of executing one action at a time. This enhances performance since it eradicates any chance of records being locked and enables a user to give numerous requests that can be handled in a sequence without any glitch.

Redis could enable the storage of sessions, compare friends of friends list, and allow one to keep a record of recent visitors.

3.318 Reef



title	Reef
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

REEF (Retainable Evaluator Execution Framework) is a scale-out computing fabric that eases the development of Big Data applications on top of resource managers such as Apache YARN and Mesos [711]. It is a Big Data system that makes it easy to implement scalable, fault-tolerant runtime environments for a range of data processing models on top of resource managers. REEF provides capabilities to run multiple heterogeneous frameworks and workflows of those efficiently. REEF contains two libraries, Wake and Tang where Wake is an event-based-programming framework inspired by Rx and SEDA and Tang is a dependency injection framework inspired by Google Guice, but designed specifically for configuring distributed systems.



title	Riak
status	10
section	NoSQL
keywords	NoSQL

As per its official documentation by Basho Technologies, Riak is a NoSQL Database which follows distributed key-value architecture [712]. Its officially named as Riak KV, however for ease of use it began to be called as Riak. It boasts of a key-value implementation that is powerful enough to store huge amounts of data of any kind, whether structured or unstructured. It even has a session tracking information storage option, which also replicates the same across the world. Riak achieves the all-round availability and faster implementation by distributing the data it saves across the world in different clusters without any large cost of operation.

A Riak ring consists of identical nodes which contain information on N, R and W, the three attributes which describe any distributed store. Here, N stores the information on replicas to be created, R and W stores information on how many replicas are needed for operations such as read and write. By storing and transmitting this logic and information to an application, Riak achieves a versatility that it can adapt to any environment settings needed for any application.

The Riak ring replicates the data into multiple other nodes, which enables easier migration of data without manual intervention. Each node shares data among the other nodes so that every node has identical data. As every node is identical, bottlenecks are avoided, however, as the number of clusters grow there is a chance of machine failure, which is detected immediately and recovers as and when the machines are brought back. Also, if any node is unavailable the data is still made accessible by getting the data from other nodes

[713].

Riak is implemented in C and Erlang and a little of Javascript and supports Python, Java, PHP and Ruby. It utilizes Rest Interface which enables it to use any client which is compatible. Riaks MapReduce implements a special case of processing, where it takes the computation source to data rather than data to the source, thus saving time in processing large computations. That is, Riak takes the code directly to the node containing the data and this reducing processing time.



title	rkt
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

rkt, formerly known as Rocket is a modern, secure, and composable application container engine developed for the production cloud-native environments. It features a pod-native approach and Linux like architecture. With its pluggable execution capabilities, it is considered a viable replacement to Docker.

"The core execution unit of rkt is the pod, a collection of one or more applications executing in a shared context" [714].

The pod in rkt is conceptually similar to pod in Kubernetes, group of containers running on the same host. The rkt gives the flexibility to apply different configuration settings using various parameters at the pod or application level. The rkt technology offers rpm or dev packages that users can install for most of the Linux distributions, and are also available to test Kubernetes integration [714].

"rkt's primary interface is a command-line tool, rkt, which does not require a long-running daemon. This architecture allows rkt to be updated in-place without affecting application containers which are currently running" [715].

There are three distinct stages of the execution within rkt. stage0 is a CLI interface that invokes the binaries and is responsible for performing a various task within the pod such as generating a Pod GUID; creating a filesystem; setting directories for stage1 and stage2;

unpacking the stage1 ACI into the pod filesystem. stage1 performs the creation of the necessary container isolation, the associated network components, and mounts to launch the pod. The final stage includes stage2, that is the environment in which the actual applications run, as launched by stage1 [715].

Unlike Docker, rkt architecture does not contain centralized daemon process and instead, it uses command lines to run the containers. This immersion helps rkt to monitor the actual container than its client process. However, one cannot run rkt commands from remote machines like docker [716].

rkt can run docker images that ultimately simplifies command line interface which makes it easier to run the Docker container on rkt with minimal migration efforts. When the core instance opened on any of the cloud provider environments, rkt converts the docker image to an application container(Apps)format and does not need docker installed to achieve this. One can install rkt on multiple Linux distributions including Ubuntu, Fedora, and Debian as a self-contained, isolated environment. With rkt installation, all security options enabled by default [716].

rkt follows an open standard for images, and this allows the open source community to develop multiple ways to build images. In the cloud world, the production environments need a scheduler to control VM's, and thankfully, both Kubernetes and Nomad support rkt. Unfortunately, neither of them have mature documentation, and decidedly fewer issues are available for troubleshooting and analysis. More recently, rkt has become a reliable alternative to Docker due to the simplified architecture, and soon one can expect

“Docker and rkt container platforms will be as good as interchangeable” [716].

3.321 Robot Operating System (ROS)



title	Robot Operating System (ROS)
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

The aptly-named Robot Operating System, or ROS, provides a framework for writing operating systems for robots. ROS offers

“a collection of tools, libraries, and conventions [meant to] simplify the task of creating complex and robust robot behavior across a wide variety of robotic platforms” [717].

ROS’ designers, the Open Source Robotics Foundation, hereinafter OSRF or the Foundation, attempt to meet the aforementioned objective by implementing ROS as a modular system. That is, ROS offers a core set of features, such as inter-process communication, that work with or without pre-existing, self-contained components for other tasks.

The OSRF designed ROS as a distributed, modular system. The OSRF maintains a subset of essential features for ROS, i.e., ROS core, to provide an extensible platform for other roboticists. The Foundation also coordinates the maintenance and distribution of a vast array of ROS add-ons, referred to as modules. ROS’ core consists of the following components: (a) communications infrastructure; (b) robot-specific features; and, (c) tools. The modules, analogous to packages in Linux repositories or libraries in other software packages such as R, provide solutions for numerous robot-related problems. General categories include (a) drivers, such as sensor and actuator interfaces; (b) platforms, for steering and image processing, etc.; (c) algorithms, for task planning and obstacle avoidance; and, (d) user interfaces,

such as tele-operation and sensor data display [718].

3.322 Rocks



title	Rocks
status	10
section	DevOps
keywords	DevOps

Rocks provides open cluster distribution solution is build targeting the scientist with less cluster experience to ease the process of deployment, managing, upgrading and scaling high performance parallel computing cluster [719]. It was initially build on linux however the latest version Rocks 6.2 Sidewinder is also available on CentOS. Rocks can help create a cluster in few days with default configuration and software packages. Rocks distribution package comes with high-performance distributed and parallel computing tools. It is used by NASA, the NSA, IBM Austin Research LAB, US Navy and many other institution for their projects.

3.323 RYA



title	RYA
status	10
section	NoSQL
keywords	NoSQL

Rya is a

“scalable system for storing and retrieving RDF data in a cluster of nodes” [720].

RDF stands for Resource Description Framework [720]. RDF is a model that facilitates the exchange of data on a network [721]. RDF utilizes a form commonly referred to as a triple, an object that consists of a subject, predicate, and object [720]. These triples are used to describe resources on the Internet [720]. Through new storage and querying techniques, Rya aims to make accessing RDF data fast and easy [722].

3.324 S4



title	S4
status	10
section	Streams
keywords	Streams

S4 is a distributed, scalable, fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of data [723]. It is built on similar concept of key-value pairs like the MapReduce. The core platform is written in Java [724]. S4 provides a runtime distributed platform that handles communication, scheduling and distribution across containers. The containers are called S4 nodes. The data is executed and processed on these S4 nodes. These S4 nodes are then deployed on S4 clusters. The user develops applications and deploys them on S4 clusters for its processing. The applications are built as a graph of Processing Elements (PEs) and Stream that interconnects the PEs. All PEs communicate asynchronously by sending events on streams. Events are dispatched to nodes according to their key in the program [723]. All nodes are symmetric with no centralized service and no single point of failure. Additionally there is no limit on the number of nodes that can be supported. [725]. In S4, both the platform and the applications are built by dependency injection, and configured through independent modules.

3.325 Saga



title	Saga
status	90
section	Interoperability
keywords	Interoperability

SAGA (Simple API for Grid Applications) provides an abstraction layer to make it easier for applications to utilize and exploit infra effectively. With infrastructure being changed continuously its becoming difficult for most applications to utilize the advances in hardware. SAGA API provides a high level abstraction of the most common Grid functions so as to be independent of the diverse and dynamic Grid environments [726]. This shall address the problem of applications developers developing an application tailored to a specific set of infrastructure. SAGA allows computer scientists to write their applications at high level just once and not to worry about low level hardware changes. SAGA provides this high level interface which has the underlying mechanisms and adapters to make the appropriate calls in an intelligent fashion so that it can work on any underlying grid system.

“SAGA was built to provide a standardized, common interface across various grid middleware systems and their versions” [727].

As SAGA is to be implemented on different types of middleware it does not specify a single security model but provides hooks to interfaces of various security models. The SAGA API provides a set of packages to implement its objectivity: SAGA supports data management, resource discovery, asynchronous notification, event generation, event delivery etc. It does so by providing set of functional packages namely SAGA file package, replica package, stream package, RPC package, etc. SAGA provides interoperability by

allowing the same application code to run on multiple grids and also communicate with applications running on others [726].

3.326 Sahara



title	Sahara
status	10
section	DevOps
keywords	DevOps

The Sahara product provides users with the capability to provision data processing frameworks (such as Hadoop, Spark and Storm) on OpenStack by specifying several parameters such as the version, cluster topology and hardware node details [728]. The solution allows for fast provisioning of data processing clusters on OpenStack for development and quality assurance and utilisation of unused computer power from a general purpose OpenStack IaaS Cloud [729]. Sahara is managed via a REST API with a User Interface available as part of OpenStack Dashboard.



title	SaltStack
status	10
section	DevOps
keywords	DevOps

The SaltStack is a management platform for centralized server infrastructure with configuration management, remote execution, monitoring and other functions, so it is just a simplified version of the puppet and enhanced version of the functions[730].

SaltStack is based on the Python language and is built with message queues (ZeroMQ) and Python third-party modules (PyZmq, PyCrypto, Jinja2, python-msgpack, and PyYAML). By deploying the SaltStack environment, users can execute commands in batches on thousands of servers, configure centralized management, distribute files, collect server data, operating system infrastructure, and package management based on different type of businesses. SaltStack is a good tool to improve work efficiency, standardize business configuration and operation[731].

SaltStack is very fast and simple to configure and maintain for different size of project and any number of servers, or for local network or a cross-data center, can manage any number of servers with diverse needs. One core features of SaltStack for parallel remote node execution commands. it uses a lot of new techniques, the network layer uses the ZeroMQ library, it also uses public and master communications, while using faster AES encryption communications, so authentication and encryption are already integrated into SaltStack. For easy extensions, the SaltStack execution routine can be written as a simple Python module and it is based on the Apache 2.0 licence and can be used for open source or proprietary projects[732].

3.328 SAML OAuth



title	SAML OAuth
status	90
section	Monitoring
keywords	Monitoring

As explained in [733], Security Assertion Markup Language (SAML) is a secured XML based communication mechanism for communicating identities between organizations. The primary use case of SAML is Internet SSO. It eliminates the need to maintain multiple authentication credentials in multiple locations. This enhances security by elimination opportunities for identity theft/Phishing. It increases application access by eliminating barriers to usage. It reduces administration time and cost by excluding the effort to maintain duplicate credentials and helpdesk calls to reset forgotten passwords. Three entities of SAML are the users, Identity Provider (IdP-Organization that maintains a directory of users and an authentication mechanism) and Service Provider (SP-Hosts the application /service). User tries to access the application by clicking on a link or through an URL on the internet. The Federated identity software running in the IdP validates the user's identity and the user is then authenticated. A specifically formatted message is then communicated to the federated identity software running at SP. SP creates a session for the user in the target application and allows the user to get direct access once it receives the authorization message from a known identity provider.

3.329 Samza



title	Samza
status	90
section	Streams
keywords	Streams

Apache Samza is an open-source near-realtime, asynchronous computational framework for stream processing developed by the Apache Software Foundation in Scala and Java [734]. Apache Samza is a distributed stream processing framework. It uses Apache Kafka for messaging, and Apache Hadoop YARN to provide fault tolerance, processor isolation, security, and resource management. Samza processes streams. A stream is composed of immutable messages of a similar type or category. Messages can be appended to a stream or read from a stream. Samza supports pluggable systems that implement the stream abstraction: in Kafka a stream is a topic, in a database we might read a stream by consuming updates from a table, in Hadoop we might tail a directory of files in HDFS. Samza is a stream processing framework. Samza provides a very simple callback-based process message API comparable to MapReduce. Samza manages snapshotting and restoration of a stream processor's state. Samza is built to handle large amounts of state (many gigabytes per partition) [735]. Whenever a machine in the cluster fails, Samza works with YARN to transparently migrate your tasks to another machine. Samza uses Kafka to guarantee that messages are processed in the order they were written to a partition, and that no messages are ever lost. Samza is partitioned and distributed at every level. Kafka provides ordered, partitioned, replayable, fault-tolerant streams. YARN provides a distributed environment for Samza containers to run in. Samza works with Apache YARN, which supports Hadoop's security model, and resource isolation through Linux CGroups [736] [734].



title	SAP HANA
status	10
section	High level Programming
keywords	High level Programming

SAP HANA[737] is an in-memory data platform that provides database, data integration and quality, in-memory OLAP, and application development services [737]. There are three main areas of functionalities in SAP HANA: Database, Analytics, and Web Application; all three areas can be integrated efficiently under SAP HANA environment to support real-time processing needs.

As a database, SAP HANA is different than traditional relational database in a sense that data is stored in memory instead of disk space, which increases processing speed tremendously by eliminating the extra step of moving data from disk to memory. If optimized, SAP HANA can support up to a petabyte of storing column-oriented, in-memory database with quick query processing power [738]. SAP HANA can also combine Online Analytical Processing (OLAP) with Online Transaction Processing (OLTP) databases to support real-time processing, data analysis and reporting [739].

In the Analytics space, SAP HANA is capable of processing spatial, graph, search, and text data and performing tasks such as prediction, streaming, time series, and machine learning [737].

Furthermore, SAP HANA is a web-based application server that hosts applications that utilize accessing, analyzing and processing data from its database. SAP HANA engine supports Node.js and JavaEE and allows the capability to support languages that is not in its native language [738].

Aside from the three key main functionalities, SAP HANA also provides administration and security tools to help monitoring processes, support continuous availability, and keep data and application secured by utilizing encryption and access management controls [740].

SAP HANA can be integrated with Hadoop and Spark, or other applications to store and access data inexpensively, it can also be implemented on popular cloud servers, such as Azure or AWS, to decrease cost of hosting and maintaining hardware and servers [739].

3.331 Sawzall



title	Sawzall
status	10
section	High level Programming
keywords	High level Programming

Google engineers created the domain-specific programming language (DSL) Sawzall as a productivity enhancement tool for Google employees. They targeted the analysis of large data sets with flat, but regular, structures spread across numerous servers. The authors designed it to handle

“simple, easily distributed computations: filtering, aggregation, extraction of statistics”,

etc. from the aforementioned data sets [741].

In general terms, a Sawzall job works as follows: multiple computers each create a Sawzall instance, perform some operation on a single record out of (potentially) petabytes of data, return the result to an aggregator function on a different computer and then shut down the Sawzall instance.

The engineer’s focus on simplicity and parallelization led to unconventional design choices. For instance, in contrast to most programming languages Sawzall operates on one data record at a time; it does not even preserve state between records [742]. Additionally, the language provides just a single primitive result function, the emit statement. The emitter returns a value from the Sawzall program to a designated virtual receptacle, generally some type of aggregator. In another example of pursuing language simplicity and parallelization, the aggregators remain separate from the formal Sawzall language (they are written in C++) because

“some of the aggregation algorithms are sophisticated and best implemented in a native language and more importantly drawing an explicit line between filtering and aggregation enables a high degree of parallelism, even though it hides the parallelism from the language itself” [741].

Important components of the Sawzall language include: szl, the binary containing the code compiler and byte-code interpreter that executes the program; the libszl library, which compiles and executes Sawzall programs

“When szl is used as part of another program, e.g. in a map-reduce program”;

the Sawzall language plugin, designated protoc_gen_szl, which generates Sawzall code when run in conjunction with Google’s own protoc protocol compiler; and libraries for intrinsic functions as well as Sawzall’s associated aggregation functionality [743].

3.332 Scalapack



title	Scalapack
status	90
section	Application and Analytics
keywords	Application and Analytics

ScaLAPACK is a library of high-performance linear algebra routines for parallel distributed memory machines. It solves dense and banded linear systems, least squares problems, eigenvalue problems, and singular value problems. It is designed for heterogeneous computing and is portable on any computer that supports Message Passing Interface or Parallel Virtual Machine [744].

ScaLAPACK is a open source software package and is available from netlib via anonymous ftp and the World Wide Web. It contains driver routines for solving standard types of problems, computational routines to perform a distinct computational task, and auxiliary routines to perform a certain subtask or common low-level computation. ScaLAPACK routines are based on block-partitioned algorithms in order to minimize the frequency of data movement between different levels of the memory hierarchy.

3.333 Scalding



title	Scalding
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

3.334 SciDB



title	SciDB
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

SciDB is an open source DBMS based on multi-dimensional array data model and runs on Linux platform [745]. The data store is optimized for mathematical operations such as linear algebra and statistical analysis. The data can be distributed across multiple nodes in a cluster.

The dimensions of the data can be either standard integers or user-defined types. Ragged arrays are also supported. The data is accessed through AQL, a SQL like language designed specifically for array operations. It supports operations such as to filter and join arrays and aggregation over the cell values. It has few similarities to Postgres in terms of user-defined scalar functions and storage manager. Old values of data are updated instead of being deleted to retain different versions of a cell. The arrays are divided into chunks and partitioned across the nodes in the cluster, with provision of caching some of them in the main memory.

3.335 scikit-learn



title	scikit-learn
status	90
section	Application and Analytics
keywords	Application and Analytics

Scikit-learn is an open source library that provides simple and efficient tools for data analysis and data mining. It is accessible to everybody and reusable in various contexts. It is built on numpy, Scipy and matplotlib and is commercially usable as it is distributed under many linux distributions [746]. Through a consistent interface, scikit-learn provides a wide range of learning algorithms. Scikits are the names given to the modules for SciPy, a fundamental library for scientific computing and as these modules provide different learning algorithms, the library is named as sciki-learn [747]. It provides an in-depth focus on code quality, performance, collaboration and documentation. Most popular models provided by scikit-learn include clustering, cross-validation, dimensionality reduction, parameter tuning, feature selection and extraction.

3.336 Security and Privacy



title	Security and Privacy
status	90
section	Monitoring
keywords	Monitoring



title	Sentry
status	10
section	Monitoring
keywords	Monitoring

Large datasets are stored on data servers. Because of its various advantages, Hadoop is used increasingly to process this data. Hadoop is used to process different databases with different formats etc. Also, depending on data, multiple users might be trying to access the data and also process it depending on their needs. As users can access various databases, they can link them and find information that is not available in single database. This can lead to users being able to find sensitive or personal information that a certain database does not want to give users access to. However, Hadoop can only control user access on file level. This works by either giving full access to file or no access at all. This model makes it difficult to implement access control as there are many databases and the overall architecture is complex. Better way of giving access to data is giving it based on users role. The way this works is by giving role based-fine grained access to users [748].

Fine grained means whether to give access to a Server, Database, Tables, Indexes or Collections. Role based means giving either of Select, Update, Query or All access to users. Scalability of this comes from how it works. Administrators can have standard tables for users, their roles and privileges given to them. Whenever a new user needs to be given access to data, they simply add them to table that works best. Same works when user access needs to be changed. This reduces chances for mistakes while giving access and reduces load on administrators from having to define access for each user individually. Also, administrator load can be further reduced by distributing task of assigning privileges to users at each database

level. This further increases control over fine-grained role-based accesses for users. Sentry is compatible with SQL query engines such as Hive, Impala and Pig [748].

3.338 Sesame



title	Sesame
status	10
section	NoSQL
keywords	NoSQL

Sesame is framework which can be used for the analysis of RDF (Resource Description Framework) data. Resource Description Framework (RDF) is a model that facilitates the interchange of data on the Web [721]. Using RFD enables us to merge data even if the underlying schemas differ. Sesame has now officially been integrated into RDF4J Eclipse project [749]. Sesame takes in the natively written code as the input and then performs a series of transformations, generating kernels for various platforms. In order to achieve this, it makes use of the feature identifier, impact predictor, source-to-source translator and the auto-tuner [750]. The feature identifier is concerned with the extraction and detection of the architectural features that are important for application performance. The impact predictor determines the performance impact of the core features extracted above. A source-to-source translator transforms the input code into a parametrized one; while the auto-tuner helps find the optimal solution for the processor.

3.339 SGE - Univa Grid Engine fa18-523-83



title	SGE - Univa Grid Engine
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

Sun Grid Engine (SGE) [751], currently known as Univa Grid Engine [752], was an open-source grid computing computer cluster developed by Sun and later bought by Oracle in 2010 [753], and then bought by Univa Corporation in 2013 [754]. SGE is the open-source version whereas Univa Grid Engine is the licensed version that is owned and supported by Univa [754]. Similar to Slurm, SGE is responsible for scheduling and managing jobs and is used in high-performance computing clusters [754].

SGE cluster consist of a head node and computational nodes. Head node usually contains a master host that runs the master daemon called `sge_qmaster` that controls the scheduling, components, and access permissions [755]. Computations nodes run execution daemon called `sge_execd` that can run three types of hosts: administration hosts, submit hosts, and execution hosts [755]. Administration hosts can run administrative activities and other tasks such as submitting, modifying, monitoring, and deleting jobs. Submits hosts limits to only able to submit and control jobs; execution hosts can only execute jobs [755].

SGE schedules jobs by allowing job submission, push them to queue and execute them with resource allocation functionality. SGE utilizes load balancer to distribute jobs and prevent resource overload from any specific nodes [756]. SGE also allows the ability to monitor jobs to check jobs statuses [www-fa18-523-83-sge-bioinformatics]. The last release of open source SGE contains additional features such as:

- Job and resource reservation
- Scheduling algorithms and topology-aware scheduling and thread binding
- GUI Installer and SGE Inspect
- Status Report and usage accounting
- Enhanced remote execution
- Multi-clustering cluster queues
- Fault-tolerance scheduling and checkpointing
- Distributed Resource Management Application API
- Job Submission Verifier
- Enhanced remote execution
- Integration with Hadoop and cloud computing service such as Amazon EC2 [754]

Before 2012, the licensed version of SGE at the time, Oracle Grid Engine, was used on the cloud for a while and could handle thousands of nodes. In 2012, it was tested for scalability on 10,000 nodes of Amazon EC2 clusters in 2012. As a result, there are still work to be done to optimize the run on 10,000 nodes and it is believed that OGE can handle all the way up to 20,000 in the future [757]. In 2018, Univa Grid Engine was able to operate on one million cores on AWS [753].



title	Shark
status	100
section	High level Programming
keywords	High level Programming

The Data Scientists when working on huge data sets try to extract meaning and interpret the data to enhance insight about the various patterns, opportunities, and possibilities that the dataset has to offer [758]. At a traditional EDW (Enterprise Data Warehouse), a simple data manipulation can be performed using SQL queries but we have to rely on other systems to apply the machine learning algorithms on these data sets. Apache Shark is a distributed query engine developed by the open source community whose goal is to provide a unified system for easy data manipulation using SQL and pushing sophisticated analysis towards the data.

“Shark is built on Hive Codebase and it has the ability to execute HIVE QL queries up to 100 times faster than Hive without making any change in the existing queries” [758].

“Shark can run both on the Standalone Mode and Cluster-Mode. Shark can answer the queries 40X faster than Apache Hive and can run machine learning algorithms 25X faster than MapReduce programs in Apache Hadoop on large data sets” [758].

Shark is a new data analysis system can process complex analytics queries on large clusters. It help in a distributed memory abstraction to provide a unified engine that can run SQL queries and sophisticated analytics functions at scale, and efficiently recovers from failures mid-query [759]. Due to this SQL queries up to 100 time

faster than Apache Hive, and machine learning programs up to 100 time faster than Hadoop by Shark. The Shark shows speedups while retaining execution engine like MapReduce and the detailed fault tolerance properties provide by such engines, unlike previous systems. It extends such an engine in several ways, including column-oriented in-memory storage and dynamic mid-query replanning, to effectively execute SQL. The result is a system that matches the speedups reported for MPP analytic databases over MapReduce, while offering fault tolerance properties and complex analytics capabilities that they lack [759].

Shark is a data Warehouse system built on top of Apache Spark which does the parallel data execution and is also capable of deep data analysis using the Resilient Distributed Datasets (RDD) memory abstraction which unifies the SQL query processing engine with analytical algorithms [758]. Based on this common abstraction, it allows running two query in the same set of workers and share intermediate data. Since RDDs are designed to scale horizontally, it is easy to add or remove nodes to accommodate more data or faster query processing. Thus, it can be scaled to the large number of nodes in a fault-tolerant manner.

Shark is a component of an open source,in-memory analytics, and fault-tolerant, distributed Spark system, which can be installed on the same cluster as Hadoop. The Shark is fully compatible with Hive and supports Hive data formats, HiveQL and user-defined functions. In addition Shark can be used to query data4 in HDFS [760], HBase [761] and Amazon S3 [762].

The best performance gains in using Impala are achieved by using the Trevni columnar storage format. In the case of Shark, their custom columnar store and compression reduced storage and query time by about 5X [762]. The Shark lets users partition tables using a specified key. In particular if tables will be joined frequently, then one can partition them using a common key. Co-partitioning is a trick used by many MPP databases to speed up joins involving massive tables [762]. RDD's are distributed objects that can be cached in-memory,

across a cluster of compute nodes. They are the fundamental data objects used in Spark. Users can create RDD's and apply machine-learning functions to them, all from within Shark. Currently machine-learning and analytic functions can be written in Scala and Java, with support for Python coming soon. Not only do users get the benefit of performing simple SQL queries and complex computations from within the same7_framework, Shark is much faster than Hadoop [762].

Thus, this new data analysis system performs query processing and complex analytics (iterative Machine learning) at scale and efficiently recovers from the failures.



title	Slurm
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

Slurm [763], also known as Slurm Workload Manager, is an open-source job schedule for Linux clusters and other Unix-like operation systems. Slurm is highly scalable, high performed, highly configurable, fault-tolerant and is easy to intergrate with other applications [764]. Slurm is currently being used by 60% of the TOP500 supercomputers [764]. Additionally, Slurm is:

“currently performing workload management on six of the ten most powerful computers in the world including the number 1 system – Tianhe-2 with 3,120,000 computing cores - as well as number 6, the GPGPU giant Piz Daint, utilizing over 5,000 NVIDIA GPGPUs” [765].

Slurm’s key functions include:

- Allocating access to users
- Providing framework that allow job scheduling and monitoring on parallel or allocated nodes
- Providing and managing job queue [766]

Slurm’s cluster controllers implement a manager daemon called slurmstld that contains a node manager, partition manager, and a job manager to allow monitoring and distribution of the jobs [767]. Each of the node implement a manager daemon called slurmd that excute and monitoring tasks on the node, also accepting commands from the slurmstld controller [767]. In addition, some of Slurm’s optional

commands/plugins that can be triggered from slurmstld or slurmd are:

- scontrol: administrative tool that helps with monitoring and DevOps
- sinfo: generate system status
- squeue: generate report status
- sacct: get jobs information
- srun: initiate jobs
- scancel: terminate jobs
- slurmdbd: record accounting information to store in database
- smap and sview: generate graphical report
- sacctmgr: database administrative tools [766]

Slurm is a flexible tool due to its capability of allowing plugins to customize functionalities based on users' needs. Some example of popular plugins are:

- **Accounting Storage**: store jobs' historical data, can be used with slurmdbd and can be integrated with other plugins such as **Account Gather Energy**, a plugin for job energy consumption gatherer, or **Job Account Gather**, a plugin for resource utilization gatherer
- **Authentication of communications**: provide authentication mechanisms
- **Cryptography**: provide digital signature
- **Scheduler**: determine how and when Slurm schedules jobs
- **Node selection**: determine resources used for job allocation [766]

Slurm is a job scheduler tool that is widely used in major supercomputer clusters due to its special architectures and features.



title	Snort
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Snort [768] is a network intrusion system capable of both detection and prevention of hazardous attacks on server systems.

Network security includes many components and practices that need to be enforced to prevent any attacks. Most of the attacks on networks happen with the intent to steal information, in many cases sensitive data that may be exploited (eg. customer payment information). One of these components being a firewall. A firewall though necessary cannot detect any harmful intrusions into the network. On the other hand, an in intrusion detection system can evaluate packets(solitary) and trigger alarms when/if it detects a hostile potential.

Drawing on Wikipedia for a quick understanding of Snort:

"Snort is a free open source network intrusion detection system (IDS)[5] created in 1998 by Martin Roesch, former founder and CTO of Sourcefire. Snort is now developed by Cisco, which purchased Sourcefire in 2013, at which Roesch is a chief security architect. Snort's open source network-based intrusion detection system (IDS) has the ability to perform real-time traffic analysis and packet logging on Internet Protocol (IP) networks. Snort performs protocol analysis, content searching and matching" [769].

Cisco defines the main benefits of snort as:

1. Rapid response: Snort allows you to write your own rules and this protect the system from any potential attacks.
2. Greater accuracy: Improvements to the Snort code is continuously brought up by the worldwide Snort Community.
3. High adaptability: Snort allows you to build upon it scode by defining your own network solutions [770].

An article by Michael Brennan(2002) elucidates that of the more promising of snort's features is its ease of configuration. It boasts of flexibility in its rules, allowing users to easily define and insert rules into its rule-base. In the case where a new attack on the system is detected, a new rule can be added into the rule-base. A second promising feature is its ability to examine packets. This means that Snort can analyze a packet to its payload and detect why it triggered the alert [771].

3.343 Solandra



title	Solandra
status	10
section	NoSQL
keywords	NoSQL

Solandra is a highly scalable real-time search engine built on Apache Solr and Apache Cassandra. Solandra simplifies maintaining a large scale search engine, something that more and more applications need. At its core, Solandra is a tight integration of Solr and Cassandra, meaning within a single JVM both Solr and Cassandra are running, and documents are stored and distributed using Cassandra's data model [772].

Solandra supports most out-of-the-box Solr functionality (search, faceting, highlights), multi-master (read/write to any node). It features replication, sharing, caching, and compaction managed by Cassandra [772].



title	Solr
status	10
section	NoSQL
keywords	NoSQL

SOLR is an open source search platform used for searching applications. It was built on top of Lucene which is a Java based full text search engine. SOLR was created by Yonik Seely in 2004 to add search capability to CNET application. It was then made an open source by Apache software foundation. SOLR can be used along with Hadoop to deal with large volumes of data. SOLR is not only a search engine but also used for storage purpose. It acts as a No-SQL database (Non-releational data storage) SOLR is scalable and ready to deploy to search and store large volumes of data. SOLR can be communicated with restful API's.

Solr enables powerful matching capabilities including phrases, wildcards, joins grouping across any datatype. This feature makes it capable for a full text search. Solr's is designed to adapt to the user needs all while simplifying configuration which makes it easily configurable. Solr is flexible and extensible. Solr publishes many well-defined extension points that make it easy to plugin both index and also query time plugins. Since it is Apache-licensed open source, you can change any code you want. Solr is a No SQL database that makes it more efficient in handling large volume of data. Since Solr is built on the battle-tested Apache Zookeeper, Solr makes it easy to scale up and down. Solr bakes in replication, distribution, rebalancing and fault tolerance out of the box. Solr takes advantage of Lucene's Near Real-Time Indexing capabilities to make sure you see your content when you want to see it. Solr publishes loads of metric data which makes monitoring simple. SOLR is built on top of Lucene. Hence, it inherits all the benefits of Lucene and also provides

ready-to-deploy service to build a search box featuring autocomplete, which Lucene doesn't provide.

3.345 Spark SQL



title	Spark SQL
status	90
section	SQL and SQL Services
keywords	SQL and SQL Services

Spark SQL is Apache Spark's module for working with structured data. Spark SQL is a new module that integrates relational processing with Spark's functional programming API [773]. It is used to seamlessly mix SQL queries with Spark programs. Spark SQL lets you query structured data inside Spark programs, using either SQL or a familiar DataFrame API. It offers much tighter integration between relational and procedural processing, through a declarative DataFrame API that integrates with procedural Spark code. Spark SQL reuses the Hive frontend and metastore, giving you full compatibility with existing Hive data, queries, and UDFs by installing it alongside Hive. Spark SQL includes a cost-based optimizer, columnar storage and code generation to make queries fast [774]. At the same time, it scales to thousands of nodes and multi hour queries using the Spark engine, which provides full mid-query fault tolerance.



title	Spark Streaming
status	10
section	Streams
keywords	Streams

Spark streaming has become increasingly popular [775] with the advent of big data with the goal of making data valuable for a company's growth. Data is streamed by batching collected live data into N time intervals based on the use case and the requirements then utilized to create final results [775].

The final result produced is also in batches. When the spark streaming is running we can view the details of the spark job in the spark console. ZeroMQ and apache Kafka are some of Spark Streaming's data sources. This can also re-launch failed tasks very easily [775].

The results are hence stored in a data store to generate report and to analyze further. Some places where the spark streaming use cases is included are:

- Uber: Uses Spark streaming to collect data from the users who use the uber mobile app for real-time analytics.
- Pinterest: Uses Spark streaming to determine how many users are sharing the pins in real time.
- Netflix: Uses Spark Streaming where billions of data received by users depending on the likes of the movie etc are collected to build real-time movie recommendations that would process.
- Yelp: determines the sentiments based on the rating and analyses that [775].

Spark Streaming is also used in:

Supply chain analytics, To give real time video experience To provide interactive experience, Real time security operations etc. Sensor data, Weather information, Fraud detection, To analyze the trend [776]

Spark streaming is currently supported in Scala, Java, and Python programming languages which typically involves the following steps:

- Initialize StreamingContext object into SparkContext and Sliding interval time.
- Specify the source of the input data
- Spark Streaming APIs define the computations
- StreamingContext processes the logic that gets defines. Using the start method.
- StreamingContext stops the streaming of the data

Spark streaming processes the real time data and provides insights by computing the log statistics [776].

3.347 Spark



title	Spark
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Apache Spark which is an open source cluster computing framework has emerged as the next generation big data processing engine surpassing Hadoop MapReduce.

“Spark engine is developed for in-memory processing as well a disk based processing. This system also provides large number of impressive high level tools such as machine learning tool M Lib, structured data processing, Spark SQL, graph processing took Graph X, stream processing engine called Spark Streaming, and Shark for fast interactive question device.”

The ability of spark to join datasets across various heterogeneous data sources is one of its prized attributes. Apache Spark is not the most suitable data analysis engine when it comes to processing (1) data streams where latency is the most crucial aspect and (2) when the available memory for processing is restricted.

“When available memory is very limited, Apache Hadoop Map Reduce may help better, considering huge performance gap.”

In cases where latency is the most crucial aspect we can get better results using Apache Storm [777].

3.348 Splunk



title	Splunk
status	90
section	Application and Analytics
keywords	Application and Analytics

Splunk is a platform for big data analytics. It is a software product that enables you to search, analyze, and visualize the machine-generated data gathered from the websites, applications, sensors, devices, and so on, that comprise your IT infrastructure or business [778]. After defining the data source, Splunk indexes the data stream and parses it into a series of individual events that you can view and search. It provides distributed search and MapReduce linearly scales search and reporting. It uses a standard API to connect directly to applications and devices. It was developed in response to the demand for comprehensible and actionable data reporting for executives outside a company's IT department [778].



title	SQL Server
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

Microsoft SQL Server is a RDBMS system by Microsoft where RDBMS stands for Relational Database Management System [779]. Its primary function is to store and retrieve data from the database as and when the software application request for it. It can be run in a single system or on a cloud to be accessed by several applications at once [779]. SQL server is being offered by Microsoft in several editions [780] which can be seen in brief as below.

1. Standard Version edition has a core engine for the database, which can be used by individuals and the entire functionality of SQL server cannot be utilized in this version. It has a very basic level of data management options and enables efficient database management.
2. Web Version edition is used as a web-based database and has very limited functionality. It is a low-cost option for users making it affordable for small applications.
3. Enterprise Version is a full version of SQL server which has the core engine service as well as the add-ons, it can support multiple users and can support memory addition during the database use. It has super-fast performance with unlimited space for virtualization and enables easier management of critical workloads.
4. Business Intelligence Version utilizes both the standard version and Business Intelligence tools like Power View,

PowerPivot and so on. It combines the power of two, that is a reasonably fast database server and high-end business intelligence tools.

5. Express Version is a free edition version of the SQL server which has the core engine and this version is scaled down but has access for multiple number of users and multiple instances of the database on a single processor. It is ideal for independent developers for building their applications as it can be installed faster and can be worked on immediately.



title	SQLite
status	10
section	SQL and SQL Services
keywords	SQL and SQL Services

3.350.1 Previous text

SQLite is a severless SQL database engine whose source code resides in the public domain [781]. SQLite databases, including tables, indices, and views, reside on a single file on the disk [781]. It has a compact library, often taking up less than KiB of space, depending on the particular configuration [781]. Performance is the tradeoff with the smaller size, i.e. performance usually runs faster when given more memory [781]. SQLite transactions comply with the ACID (Atomicity, Consistency, Isolation, Durability) properties [781] [782]. SQLite does not require administration or configuration [783]. There are some limitations associated with SQLite, such as the inability to perform Right Outer Joins, read-only views, and access permissions (other than those that are associated with regular file acces permissions) SQLite does not compare directly with clien/server databases such as MySQL as they are both trying to solve different problems [784] [783]. While database engines such as MySQL aim to provide a shared database, with different access permissions to different individuals/applications, SQLite has the goal of being a local repository of data for applications [784]. While SQLite is not appropriate for every situation, there certainly exists situations where it can prove to be a prudent choice for data management needs [784].

3.350.2 New

SQLite [781] is an open-source transactional database engine that is

widely distributed and used throughout the world in many applications. It was created in May 2000 as part of a project to design a database that does not rely on a database management system nor a database administrator. SQLite does not contain a server component, and works very well as an embedded component within particular applications such as web browsers, operating systems and mobile phones. Google Chrome, Safari, and Android browser are just a few examples of web browsers that leverage SQLite as an embedded database platform with the application [785].

SQLite uses the standard SQL syntax within a standalone command prompt shell. Users have the ability to create, update, and delete tables as well as insert new records within tables. Users can also design and run queries similarly to other database management systems with the only main advantage being that the tool is completely free to download. One of the main drawbacks is the inability for SQLite to perform right outer joins, create read-only views, and to work with access permissions other than those associated with typical file access.

The SQLite platform is available in 32 and 64 bit installations, and is capable of handling up to 140 terabytes of data. All of the SQLite objects can reside on a single disk, including the library which is very compact in size [781]. The platform is supported by a large user community and a robust support team complete with very detailed documentation.

"SQLite database files are recommended by the US Library of Congress as the storage format for long-term preservation of digital content" [786].

SQLite can be downloaded directly from the SQLite.org website. The website contains precompiled binary install files for a variety of operating systems including Windows, Linux, and Android. The download site also contains a comprehensive set of documentation to assist in the download, installation and setup of the tool.

"The SQLite source code is maintained in three

geographically-dispersed self-synchronized Fossil repositories that are available for anonymous read only access" [787].

SQLite also provides bindings for several programming languages related to data science such as Python, R, and MATLAB [785].

3.351 Sqoop



title	Sqoop
status	90
section	Data Transport
keywords	Data Transport

Apache Sqoop is a tool to transfer large amounts of data between Apache Hadoop and sql databases [788]. The name is a Portmanteau of SQL + Hadoop. It is a command line interface application which supports incremental loads of complete tables, free form (custom) SQL Queries and allows the use of saved and scheduled jobs to import latest updates made since the last import. The imports can also be used to populate tables in Hive or Hbase. Sqoop has the option of export, which allows data to be transferred from Hadoop into a relational database. Sqoop is supported in many different business integration suits like Informatica Big Data Management, Pentaho Data Integration, Microsoft BI Suite and Couchbase [789].

3.352 Sqrrl



title	Sqrrl
status	10
section	NoSQL
keywords	NoSQL

3.353 SSH



title	SSH
status	10
section	Data Transport
keywords	Data Transport

SSH is a cryptographic network protocol to provide a secure channel between two clients over an unsecured network [790]. It uses public-key cryptography for authenticating the remote machine and the user. The public-private key pairs could be generated automatically to encrypt the network connection. ssh-keygen utility could be used to generate the keys manually. The public key then could be placed on the all the computers to which the access is required by the owner of the private key. SSH runs on the client-server model where a server listens for incoming ssh connection requests. It is generally used for remote login and command execution. Its other important uses include tunneling (required in cloud computing) and file transfer (SFTP). OpenSSH is an open source implementation of network utilities based on SSH [791].

3.354 Stackato



title	Stackat
status	10
section	Application Hosting Frameworks
keywords	Application Hosting Frameworks

Hewlett Packard Enterprise or HPE Helion Stackato is a platform as a service (PaaS) cloud computing solution. The platform facilitates deployment of the user's application in the cloud and will function on top of an Infrastructure as a service (IaaS) [792]. Multiple cloud development is supported across AWS, vSphere, and Helion Openstack. The platform supports the following programming languages: native .NET support, java, Node.js, python, and ruby. This flexibility is advantageous compared to early PaaS solutions which would force the customer into utilizing a single stack. Additionally, this solution has the capacity to support private, public and hybrid clouds [793]. This capability user has to not have to make choices of flexibility over security of sensitive data when choosing a cloud computing platform.

3.355 Stomp



title	Stomp
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

3.356 Storm



title	Storm
status	90
section	Streams
keywords	Streams

Apache Storm is an open source distributed computing framework for analyzing big data in real time [794]. refers storm as the Hadoop of real time data. Storm operates by reading real time input data from one end and passes it through a sequence of processing units delivering output at the other end. The basic element of Storm is called topology. A topology consists of many other elements interconnected in a sequential fashion. Storm allows us to define and submit topologies written in any programming language.

Once under execution, a storm topology runs indefinitely unless killed explicitly. The key elements in a topology are the spout and the bolt. A spout is a source of input which can read data from various datasources and passes it to a bolt. A bolt is the actual processing unit that processes data and produces a new output stream. An output stream from a bolt can be given as an input to another bolt [795].

3.357 Apache Flink



title	Apache Flink
status	90
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Apache Flink is an open-source stream processing framework for distributed, high-performing, always-available, and accurate data streaming applications. Apache Flink is used in big data application primarily involving analysis of data stored in Hadoop clusters. It also supports a combination of in-memory and disk-based processing as well as handles both batch and stream processing jobs, with data streaming the default implementation and batch jobs running as special-case versions of streaming application [796].

3.358 Summingbird



title	Summingbird
status	10
section	High level Programming
keywords	High level Programming

Summingbird can be described as

“a library that lets you write MapReduce programs that look like native Scala or Java collection transformations and execute them on a number of well-known distributed MapReduce platforms, including Storm and Scalding” [797].

Summingbird is open-source and is a domain-specific Scala implemented language [798]. It combines online and batch MapReduce computations into one framework [798]. It utilizes the platforms Hadoop for batch and Storm for online process execution [798]. The open-source Hadoop implementation of MapReduce is a tool which those responsible for data management use to handle problems related to big data [798]. Summingbird uses an algebraic structure called a commutative semigroup to perform aggregations of both batch and online processes [798]. A commutative semigroup is a particular type of semigroup

“where the associated binary operation is also commutative” [798].

The types of data that Summingbird takes as inputs are streams and snapshots [798]. The types of data Summingbird jobs generate are called stores and sinks [798]. Stores are

“an abstract model of a key-value store”

while sinks are unaggregated tuples from a producer [798]. Summingbird aims to simplify the process of both batch and online analytics by exploiting

“the formal properties of algebraic structures”

to integrate the various modes of distributed processing [798].

3.359 Swift



title	Swift
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Swift is a general-purpose, multi-paradigm, compiled programming language. It has been developed by Apple Inc. for iOS, macOS, watchOS, tvOS, and Linux. This programming language is intended to be more robust and resilient to erroneous code than Objective-C, and more concise. It has been built with the LLVM compiler framework included in Xcode 6 and later and, on platforms other than Linux. C, Objective-C, C++ and Swift code can be run within one program as Swift uses the Objective-C runtime library [799].

Swift supports the core concepts that made Objective-C flexible, notably dynamic dispatch, widespread late binding, extensible programming and similar features. Swift features have well-known safety and performance trade-offs. A system that helps address common programming errors like null pointers was introduced to enhance safety. Apple has invested considerable effort in aggressive optimization that can flatten out method calls and accessors to eliminate this overhead to handle performance issues.

3.360 Tableau



title	Tableau
status	10
section	Application and Analytics
keywords	Application and Analytics

Tableau is a family of interactive data visualization products focused on business intelligence [800]. The different products which tableau has built are: Tableau Desktop, for individual use; Tableau Server for collaboration in an organization; Tableau Online, for Business Intelligence in the Cloud; Tableau Reader, for reading files saved in Tableau Desktop; Tableau Public, for journalists or anyone to publish interactive data online. [801]. Tableau uses VizQL as a visual query language for translating drag-and-drop actions into data queries and later expressing the data visually. Tableau also benefits from an Advanced In-Memory Technology for handling large amounts of data. The strengths of Tableau are mainly the ease of use and speed. However, it has a number of limitations, which the most prominent are unfitness for broad business and technical user, being closed-source, no predictive analytical capabilities and no support for expanded analytics.



o: this has lots of advertisement terms in it, can you quantify the advertisement such as why is it robust and superior and so on

title	Tajo
status	10
section	High level Programming
keywords	High level Programming

Apache Tajo is a robust big data relational and distributed data warehouse system for Apache Hadoop [802]. Apache Tajo was designed to handle ad-hoc queries, online aggregation, and extract-transform-load processing on very large data sets. This system was designed for querying and analyzing data sets. The data-sets that are utilized by Apache Tajo are stored using Hadoop's Distributed File System also known as HDFS. On top of this Apache Tajo has its own query engine instead of utilizing the MapReduce framework like other Hadoop systems.

Apache Tajo allows for superior data querying by supporting SQL on Hadoop. This way users can interact with data on Hadoop via easy SQL commands [803]. Apache Tajo provides users with a SQL shell to query the HDFS data sets. Clusters are utilized on this framework to carry out queries. Clusters are made up of a master node and several workers. The master node works to plan and organize the workers to divide each query into small tasks that can be accomplished in pieces by each worker. This allows for more flexible and powerful querying ability than what is typically seen on Hadoop via the MapReduce framework. Some of the SQL functions that are available in the Apache Tajo framework are Math functions, DateTime Functions, JSON functions, and string functions [803].

Apache Tajo supports several other data formats including JSON, Text File, Parquet, Sequence File, Protocol Buffer, etc. While Apache Tajo supports several different data formats, it also allows the system to share data across different repositories. The goal of this is to allow for online analytical processing of (OLAP) data. Using this process allows Apache Tajo to scale up and distribute the load of querying the data. One of the major features of Apache Tajo is its low latency, scalability and ease of data flow. The latency, scalability, and formats that Tajo is able to support is a reason why Tajo is such a powerful system.

3.362 Talend



title	Talend
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Talend is Apache Software Foundation sponsor Big data integration tool design to ease the development and integration and management of big data, Talend provides well optimized auto generated code to load transform, enrich and cleanse data inside Hadoop, where one don't need to learn write and maintain Hadoop and spark code. The product has 900+ build-in components feature data integration

Talend features multiple products that simplify the digital transformation tools such as Big data integration, Data integration, Data Quality, Data Preparation, Cloud Integration, Application Integration, Master Data management, Metadata Manager. Talend Integration cloud is secure and managed integration Platform-as-a-service (iPaaS), for connecting, cleansing and sharing cloud on premise data.



title	Taverna
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Apache Taverna [804] is a tool for developing and implementing workflows. It is an open-source software, initially called Taverna Workbench, is now a project under the Apache Incubator project. One of the key features of Taverna is that it allows developers to include many different elements together including REST web services and SOAP [805].

When starting a scientific experiment, the main emphasis is not just the result obtained, but the procedure or method followed. Scientific workflows have thus popped up in numerous applications and experiments to reduce difficulties. Apache Taverna was introduced initially under the myGrid project. In the words of its developers,

"They provide a high-level declarative way of specifying what a particular *in silico* experiment modelled by a workflow is set to achieve, not how it will be executed" [806].

Experimental procedures can be automated by Taverna, integrating numerous remote and local services, including chemistry, physics, biology, music, meteorology, among many others. To sum up Taverna's many uses, it allows the experimenter to conduct complex analytics on data. All this can be achieved from a simple PC (windows or apple or Unix OS) with computational resources that can be both public and private. Taverna is mostly used by scientists who have limited access to resources and support.

Taverna's numerous features include workflow repository, service catalog, workbench, workflow components, web portals or gateways, client/user interfaces, command line, program APIs, being the more frequently used.

"Taverna performs multi-step or repetitive analysis that involves invoking several services, enables users to copy and paste results of various services, and automates processes. The workflow provenance gives a detailed trace of workflow execution such as services executed, when, which inputs were used and what outputs were produced. Currently, Taverna works with Biomart, SoapLab, SADI, R and Bioconductor" [807].

According to an article published in the Nucleic Acids Research Journal by Wolstencroft, Haines and others (2013), it is difficult to access and integrate resources that are distributed over many services, especially if they change over time and are incompatible. In this regard, the Taverna Tool Suite enables the available distribute data and analysis methods by supplementing the workflow with service discovery [808].



title	TensorFlow
status	10
section	Application and Analytics
keywords	Application and Analytics

TensorFlow [809] is a software library that utilize dataflow that is most commonly used for training model especially in Deep Learning and neural networks which evolves around mathematical computation. Computation is done through the dataflow structure where Tensors, the data itself, are being passed to nodes, which perform mathematical computation, and each nodes are connected by edges, indicate the flow of the Tensors [810].

Computation in Tensorflow is being mapped into different cluster of machines which are not limited to only desktop machines and server, but also different CPUs and GPUs of mobile devices [811]. As such, Tensorflow APIs are available in multiple programming language such as Python, Javascript, C++, Java, Go, and Swift. This framework, dataflow, is most commonly used especially for parallel computing that gives certain advantages: 1) distribution execution, allowing Tensorflow to distribute loads after partition to different machines, 2) compilation, which increases the performance speed by combining multiple operations of the same flow, and 3) portability, which allows the program to run on multiple languages [812].

Tensorflow, with a strong training model framework, has been utilized in variety of use case by developers coming from multiple field of interests. For instance, the use image recognition in health industry which utilize Deep Learning to learn about retina of diabetes patients. Doctors then use this model to predict the likelihood of patients who have similar retina orientation whether they have diabetes. Another use case is in biology field where scientist keep

track of almost extinct species using drones to capture millions of videos and images in large area. Therefore, the automation that TensorFlow provide allows developers to scale their research utilizing its Machine Learning [813].



title	Terraform
status	10
section	DevOps
keywords	DevOps

The word Terraform means in the real world to transform a planet so as to resemble earth, so that it can support human life. Terraform by Hashicorp is also used in the same sense. It is used to improve and change the existing infrastructure, destroy the ones which are no longer used and also create new ones [814]. It is one of the most famous open source infrastructure automation tools. Using Terraform one can define a datacenter in any configuration language, which can be leveraged to build any infrastructure or use any cloud service provider such as AWS, Azure. It is primarily written in Go and can be used on any operating system such as Linux, MacOS, Microsoft Windows or Solaris. The following features make Terraform as good as it is: It describes the infrastructure in a high-level configuration level which makes it extremely effective and reusable. Using this approach, there is a marked increase in the productivity and considerable decrease in human error. It can be used to provision resources which are available on any other infrastructure provider. It allows developers to choose and pick the infrastructure that is best suited for the application they are running. They are free to provision their own resources without being bothered about Terraform configurations. Also, the configurations can be stored as version controls which can be shared and collaborated with other members of the team [815]. Currently Terraform has more than 125 infrastructure providers and 1000+ resources. The infrastructure may be defined in HCL terraform syntax or JSON format. The latest release of Terraform 0.11.0 includes massive improvements to the registry integration and CLI workflow. There are also large number of improvements in a number of major providers. Terraform is currently

being used by Uber, Instacart, HotelTonight, Starbucks etc to orchestrate their infrastructure [816].

3.366 Tez



title	Tez
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Apache Tez is open source distributed execution framework build for writing native YARN application. It provides architecture which allows user to convert complex computation as dataflow graphs and the distributed engine to handle the directed acyclic graph for processing large amount of data. It is highly customizable and pluggable so that it can be used as a platform for various application. It is used by the Apache Hive, Pig as execution engine to increase the performance of map reduce functionality [817]. Tez focuses on running application efficiently on Hadoop cluster leaving the end user to concentrate only on its business logic. Tez provides features like distributed parallel execution on hadoop cluster, horizontal scalability, resource elasticity, shared library reusable components and security features. Tez provides capability to naturally map the algorithm into the hadoop cluster execution engine and it also provides the interface for interaction with different data sources and configurations.

Tez is client side application and just needs Tez client to be pointed to Tez jar libraries path makes it easy and quick to deploy. User can have multiple tez version running concurrently. Tez provides DAG API's which lets user define structure for the computation and Runtime API's which contain the logic or code that needs to be executed in each transformation or task.

3.367 Theano



title	Thean
status	90
section	Application and Analytics
keywords	Application and Analytics

Theano is a Python library. It was written at the LISA lab. Initially it was created with the purpose to support efficient development of machine learning (ML) algorithms. Theano uses recent GPUs for higher speed. It is used to evaluate mathematical expressions and especially those mathematical expressions that include multi-dimensional arrays. Theano's working is dependent on combining aspects of a computer algebra system and an optimizing compiler. This combination of computer algebra system with optimized compilation is highly beneficial for the tasks which involves complicated mathematical expressions and that need to be evaluated repeatedly as evaluation speed is highly critical in such cases. It can also be used to generate customized C code for number of mathematical operations. For cases where many different expressions are there and each of them is evaluated just once, Theano can minimize the amount of compilation and analyses overhead [818].

3.368 three.js



title	three.js
status	90
section	Application and Analytics
keywords	Application and Analytics

Three.js is an API library with about 650 contributions till date, where users can create and display an animated 3D computer graphics in a web browser. It is written in javascript and uses WebGL, HTML5 or SVG. Users can animate HTML elements using CSS3 or even import models from 3D modelling apps [819]. In order to display anything using three.js we need three basic features, which are scene, camera and renderer. This will result in rendering the scene with a camera. In addition to these three features, we can add animation, lights (ambience, spot lights, shadows), objects (lines, ribbons, particles), geometry etc [820].



title	Thrift
status	10
section	Message and Data Protocols
keywords	Message and Data Protocols

Apache Thrift [821] is a software framework that is equipped with code generation engine for scalable cross-language services. Thrift works between programming languages: C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, JavaScript, Node.js, Smalltalk, Ocaml and Delphi and other languages. If we need multiple programming language support, schema definitions are required like Thrift and Protocol Buffers. The schema definitions are used to map them to a target programming language. Thrift supports many programming languages because there are many contributed developers [822].

Thrift was first developed by Facebook, and open sourced in April 2007, then became a Apache Incubator project in May 2008. Thrift uses Interface Definition Language (IDL) to allow the definition of data types, and generate codes for RPC clients [821].

“Specifically, Thrift allows developers to define data types and service interfaces in a single language-neutral file and generate all necessary code to build RPC clients and servers” [823].

The process of creating a service starts with service design using Thrift documentation written in Interface Definition Language(IDL). The defined data types will translate into different types in different programming languages. Then using the thrift documentation as an input, Apache Thrift Compiler generates code for RPC clients and the server to communicate in different programming languages

seamlessly [821]. Apache Thrift is used by many companies and projects. According to the official Apache Thrift website, companies using Apache Thrift in their production services are Cloudera, Evernote, Facebook, last.fm, Mendeley, OpenX, Pinterest, Quora, RapLeaf, reCaptcha, Siemens, Uber. The open source projects using Apache Thrift are Microsoft Robust Distributed System Nucleus (rDSN), Twitter Finagle, Twitter Scrooge. The other Apache projects that are also using Thrift are Aurora, Hadoop, HBase, Parquet, Storm [821].

3.370 Tika



title	Tika
status	90
section	Extraction Tools
keywords	Extraction Tools

"The Apache Tika toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more [824]."

3.371 TinkerPop



title	TinkerPop
status	90
section	Application and Analytics
keywords	Application and Analytics

ThinkerPop is a graph computing framework from Apache software foundation [825]. Before coming under the Apache project, ThinkerPop was a stack of technologies like Blueprint, Pipes, Frames, Rexters, Furnace and Gremlin where each part was supporting graph-based application development. Now all parts are come under single ThinkerPop project repo [826]. It uses Gremlin, a graph traversal machine and language. It allows user to write complex queries (traversal), that can use for real-time transactional (OLTP) queries, graph analytic system (OLAP) or combination of both as in hybrid. Gremlin is written in java [827]. ThinkerPop has an ability to create a graph in any size or complexity. Gremlin engine allows user to write graph traversal in Gremlin language, Python, JavaScript, Scala, Go, SQL and SPARQL. It is capable to adhere with small graph which requires a single machine or massive graphs that can only be possible with large cluster of machines, without changing the code.



title	Titan:db
status	10
section	NoSQL
keywords	NoSQL

Titan [828] is a graph database that can be optimized for storage and query of graphs that can contain hundreds of billions of edges and vertices that are spread across multiple machine clusters. This multiple machine cluster that can support many concurrent users in real time. Its main integration platform is based on Apache and is open sourced. Titan specifically sits upon the Apache Cassandra database. Titan is a beneficial tool because it can access storage and other computational methods that normally one machine is unable to provide. Titan can be described as

"a graph database engine that integrates existing solutions as building blocks to form a system" ???.

The primary language that users use to traverse their graphs is Gremlin. Gremlin is an Apache query language that provides ease of transport through large datasets. Titan is elastic enough to provide the introduction and removal of different machines. Titan is different from other database technologies since it does not require the use of master-slave read/write orientations. Titan also uses Hadoop for batch graph analysis and processing. Aside from Apache Cassandra, Titan also uses HBase and BerkeleyDB database storage systems [828].

Titan holds applications that act in two specific ways. The first is the embedding of Titan inside applications provided by the Gremlin language inside the same Java Virtual Machine. Storage may be remote or local when database handling is provided within the same

Java Virtual Machine. The second method is used by interaction with local or remote Titan instances from query additions along the same server.

3.373 Torch



title	Torch
status	90
section	Application and Analytics
keywords	Application and Analytics

Torch is a open source machine learning library, a scientific computing framework [829]. It implements LuajIT programming language and implements C/CUDA. It implements N-dimensional array. It does routines of indexing, slicing, transposing etc. It has an interface to C language via scripting language LuajIT. It supports different artificial intelligence models like neural network and energy based models. It is compatible with GPU. The core package of is torch. It provides a flexible N dimensional array which supports basic routings. It has been used to build hardware implementation for data flows like those found in neural networks.

3.374 Torque



title	Torque
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

3.375 TOSCA



title	TOSCA
status	90
section	Interoperability
keywords	Interoperability

3.376 Totem



title	Totem
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

Totem is a project to overcome the current challenges in graph algorithms. The project is research the Networked Systems Laboratory (NetSysLab) The issue resides in the scale of real world graphs and the inability to process them on platforms other than a supercomputer. Totem is based on a bulk synchronous parallel (BSP) model that can enable hybrid CPU/GPU systems to process graph based applications in a cost effective manner [830].

3.377 Triana



title	Triana
status	90
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Triana is an open source problem solving software that comes with powerful data analysis tools [831]. Having been developed at Cardiff University, it has a good and easy-to-understand User Interface and is typically used for signal, text and image processing. Although it has its own set of analysis tools, it can also easily be integrated with custom tools. Some of the already available toolkits include signal-analysis toolkit, an image-manipulation toolkit, etc. Besides, it also checks the data types and reports the usage of any incompatible tools. It also reports errors, if any, as well as useful debug messages in order to resolve them. It also helps track serious bugs, so that the program does not crash. It has two modes of representing the data - a text-editor window or a graph-display window. The graph-display window has the added advantage of being able to zoom in on particular features. Triana is specially useful for automating the repetitive tasks, like finding-and-replacing a character or a string.

3.378 Trident



title	Trident
status	10
section	Workflow-Orchestration
keywords	Workflow-Orchestration

Apache Trident is a

“high-level abstraction for doing realtime computing on top of Apache Storm” [832].

Similarly to Apache Storm, Apache Trident was developed by Twitter. Furthermore, Trident as a tool

“allows you to seamlessly intermix high throughput (millions of messages per second), stateful stream processing with low latency distributed querying” [832].

The five kinds of operations in Trident are described as

“Operations that apply locally to each partition and cause no network transfer, repartitioning operations that repartition a stream but otherwise don’t change the contents (involves network transfer), aggregation operations that do network transfer as part of the operation, operations on grouped streams and “merges and joins” [833].

These five kinds of operations (i.e. joins, aggregations, grouping, functions, and filters) and the general concepts of Apache Trident are described as similar to

“high level batch processing tools like Pig or Cascading” [832].



title	Twister
status	10
section	Basic Programming Model and Runtime, SPMD, MapReduce
keywords	Basic Programming Model and Runtime, SPMD, MapReduce

The Twister framework is used to perform iterative map reduce function using the publish/subscribe messaging infrastructure [834]. Map reduce tasks once configured can be used many times and manages a lot of data. The programming extensions given to map reduce like “broadcast” and “Scatter type” improves the efficiency. Twister is predominantly used for big data batch processing. The twister architecture is very flexible. It reads data from the local disk and handles the intermediate data in the distributed memory of the worker node [834].

The config phase introduced by Twister loads any static data that is required for both map. For running a Map/Reduce task, loading static data for once is also helpful. The messaging infrastructure responsible for data transfer is called a broker network. To add heavy computational weight, Twister uses a Fat map task on the map side[835].

“Twister programming model does not guarantee the availability of the state information in map/reduce tasks across invocations” [www-Ekanayakeetal2010twister].

Along with map reduce twister comes with a new phase that adds up the output coming from all the reducer called combine operation [www-Ekanayakeetal2010twister].

Programming extensions are added to the map reduce in twister. Twister uses an iterative functionality like `mapReduceBCast(Value value)` where a single value is sent to all map tasks. In addition, map/reduces task can be configured from a set of value. Eg: '`configureMaps(Value[]values)`' and '`configureReduce(Value[]values)`' where value can be a set of parameter or a block of data are two extensions that is provided by twister [835].

To support map-reduce features, twister provides:

- Light weight java code
- Efficient tools for data management
- Good support for interactive map-reduce to perform faster computations.
- Enhance map-reduce run time
- When fault tolerance is enabled, it automatically recovers from the fault
- Accessing the Data via local disks etc [835].

Twister has three main entity:

- Client-Side Driver
- Twister Daemon
- The broker Network [835].

To pass data directly, twister keeps all the data read as native file. Additionally, they perform operations like:

- Directory creation
- Directory deletion
- Input files copied and distributed across worker nodes
- Output files collected and transferred to a given location.
- Partition files created and distributed for a set of data. [www-Ekanayakeetal2010twister]

Twister runtime efficiency is increased using subscribe messaging infrastructure. The communication network can be made fault tolerant independent of twister runtime [www-

Ekanayakeetal2010twister].



title	Twitter Heron
status	10
section	Streams
keywords	Streams

Twitter Heron [836] is defined as

“a real-time analytics platform that is fully API-compatible with Strom” [837].

Strom is the distributed stream computation open-sourced system previously used by Twitter. When the needs for data processing has increased, Twitter decides to build a new system, which is heron now, instead of spending a lot of time extending storm. The basic function architecture of Heron is that

there is a scheduler receiving topology submission from users and the scheduler would

“runs each topology as a job consisting of several containers. One of the containers runs the topology master, responsible for managing the topology. The remaining containers each run a stream manager responsible for data routing, a metrics manager that collects and reports various metrics and a number of processes called Heron instances which run the user-defined spout/bolt code” [837].

To handle the great amount of data from Twitter, Heron has a back pressure mechanism. This mechanism works like a valve of a pipe. It dynamically detects the data flow, according to which it adjusts the data flow rate in a topology to make sure the data accuracy is kept at the same level [837]. Also, tasks would be run in isolation of process-

level, which means the performance could be understood easily to reduce the difficulty of debugging [837].

To express the number of instances of tweet, mechanism call parallelism is used in Heron. Parallelism stands for the structure that nodes of topology have number noted beside them, where those numbers are noted based on incoming and outgoing data rate by developers, to reveal the number of instances needed to run on CPUs in parallel [838].

Furthermore, the execution of topology are free to coordinate. By default, instances would be run in a single container, but developers could specify how many containers should be used for running these instances, plus the CPU, disk space and memory [838]. Once the settings are finished and the instances are packed into containers, the scheduler would run those containers properly [838].



title	Tycoon
status	10
section	NoSQL
keywords	NoSQL

Tycoon [839] provides software, application, and networking services to mobile and wireless applications. Most of their web solutions in place use industry standards such as Ajax, PHP, EJB, and more. Tycoon is placed heavily in both banking and retail sectors. Tycoon is a network server on top of Kyoto Cabinet's key-value storage database. Its purpose is to aid in concurrency of data access. Its keyfeatures include master-slave and master-master replication, databases that are memory-stored, databases with hash and tree-based formats, and server-side scripting in the Lua API. Tycoon uses its own HTTP binary protocol to increase its performance. Storage libraries can be written in many languages, but the most common is Python. It also supports memcached protocol, which is another database-caching system. This is to be used if a user wishes to update a memory space that is larger than normal. Like memcached, Tycoon also has an auto expiration mechanism. The Tycoon server is able to handle around 10,000 connections at the same time. The main server is written in C++ and is available on platforms that have corresponding API library extensions.

The architecture is dynamically designed to increase speed in performance in data pulling scenarios. For example, clients have direct access to the web server which is able to access memcached protocols. From there, the Tycoon base is considered in a slave state that can then access the master Tycoon database which is also able to move sideways and pull from other master Tycoon systems [839].



title	Tyrant
status	10
section	NoSQL
keywords	NoSQL

3.383 old text

Tyrant provides network interfaces to the database management system called Tokyo Cabinet. Tyrant is also called as Tokyo Tyrant. Tyrant is implemented in C and it provides APIs for Perl, Ruby and C. Tyrant provides high performance and concurrent access to Tokyo Cabinet. The results of performance experiments between Tyrant and Memcached and MySQL can be viewed in the blog [840].

Tyrant was written and maintained by FAL Labs [841]. However, according to FAL Labs, their latest product Kyoto Tycoon is more powerful and convenient server than Tokyo Tyrant [842].

3.384 New text

Tokyo Tyrant is comprised of several packages of network interfaces that link to a complex database management system entitled Tokyo Cabinet [841]. The Tokyo Cabinet is a set of routines used for the management of key-value databases, and was initially sponsored by the Japanese social media site Mixi [843]. Tokyo Tyrant provides a variety of methods to connect to the Tokyo cabinet database manager. The application includes a process whereby allowing for effective database management as well as its access library for client base applications [841].

Below is some additional technical information from the fallabs website:

"The server features high concurrency due to thread-pool modeled implementation and the epoll/kqueue mechanism of the modern Linux/*BSD kernel. The server and its clients communicate with each other by simple binary protocol on TCP/IP. Protocols compatible with memcached and HTTP are also supported so that almost all principal platforms and programming languages can use Tokyo Tyrant. High availability and high integrity are also featured due to such mechanisms as hot backup, update logging, and replication. The server can embed Lua, a lightweight script language so that you can define arbitrary operations of the database" [844].

You must install Tokyo Cabinet prior to using Tokyo Tyrant. The Tokyo Tyrant application is extracted into a new working directory location whereby a configuration script can be executed to complete the installation. The ttserver command is vital to the operation of the server, as well as determining the best usage of memory for the application.

"The command ttserver runs the server managing a database instance. Because the database is treated by the abstract API of Tokyo Cabinet, you can choose the scheme on start-up of the server. Supported schema are on-memory hash database, on-memory tree database, hash database, and B+ tree database. This command is used in the following format. `dbname' specifies the database name. If it is omitted, on-memory hash database is specified" [844].

A new, more robust version of Tokyo Cabinet entitled Kyoto Cabinet has been released and has taken the place of the original Tokyo Cabinet platform. The Kyoto Cabinet exhibits extraordinarily fast performance using a smaller footprint, and is very scaleable. Kyoto Cabinet was developed using the C++ language and can be leveraged as an API within the Java, Python, Ruby, Perl, and Lua. [483]. A newly released database server, Kyoto Tycoon, works in tandem with Kyoto Cabinet to manage the intricate web of network connections and server processes for the applications for which it supports. Kyoto Tycoon is capable of managing concurrent network connections to Kyoto Cabinet and functions primarily as a remote network interface. Kyoto Cabinet, used in conjunction with Kyoto Tycoon, is the next step up for the Tokyo Tyrant platform and is touted as being far superior and highly recommended as a necessary replacement. Fal Labs, the architects of Tokyo Tyrant strongly recommend upgrading in order to take advantage of the new features of Kyoto Cabinet and Kyoto Tycoon [841].



title	Ubuntu MaaS
status	90
section	DevOps
keywords	DevOps

Ubuntu MaaS (Metal as a Service) allows a user to deploy and manage physical hardware. Similar to IaaS (Infrastructure as a Service) applications Ubuntu MaaS provides the capabilities to configure nodes, install a preferred OS and install applications [845]. The difference is in the level of detail the user has access to. IaaS typically refers to the configuration of virtual machines, meaning that the user does not have access to the actual infrastructure [846]. With Ubuntu MaaS data center managers have the capability to configure key details of individual servers for optimal performance. [847]

Ubuntu MaaS works by utilizing a tiered architecture of region nodes and rack nodes that are used to manage the hardware [848]. The region controller deals with requests and works with the rack controller. The rack controller provides local services and also caches the operating system image [848]. Ubuntu MaaS also provides high availability at the database level. To achieve this the region controller automatically switches gateways in the event of a rack controller failure [848]. In addition, Ubuntu MaaS allows grouping of machines to create physical availability zones. This allows the service to scale and helps streamline the deployment of complex solutions [848].

Ubuntu MaaS manages the implementation of a new node through a process called the node lifecycle [848]. The first step is to bring the hardware into the ecosystem and to take an inventory of the physical components at its disposal. Next, the set up phase gets the server ready for deployment. Once the set up phase is complete the server can be deployed or retired.

“MAAS is designed for devops at scale, in places where bare metal is the best way to run your app. Big data, private cloud, PAAS and HPC all thrive on MAAS.” [847]

Ubuntu MaaS was designed for situations where the user is managing a physical data center and where flexibility, control and efficiency are important. IaaS is similar to MaaS and is ideal for setting up virtual machines in a cloud environment. Some had questions about the need for Ubuntu MaaS due to these similarities [849]. However, as the product developed it has become a useful tool for data center managers.



title	UIMA
status	10
section	Extraction Tools
keywords	Extraction Tools

UIMA stands for unstructured information management architecture and was originally developed by IBM [850]. However currently a reference implementation of UIMA called Apache UIMA is widely used. UIMA is used for AI related task such as Natural Language Processing, Information Retrieval and Machine learning as it provides us a platform to use and analyze unstructured data such as videos, audios, images etc. It creates a relation between the unstructured data and structured world so that we can make more sense out of the data. The unique feature of UIMA architecture is that the applications can be decomposed into various components each performing different analytic operations mentioned above and the framework manages the dataflow between these components and also to the database which stores the unstructured data [851]. This promotes the reusability of components and reduces the duplicity of analytic operations. These components are written in Java with C++ enablement and can be installed on any platform or operating system. One of the biggest usages of UIMA comes in the field of text search. Large amount of textual unstructured data needs to be stored, processed and analyzed for the text to be made searchable. Various NLP techniques like tokenization, lemmatization, named entity detection, relationship detection need to be employed on the text data, all of which can be handled by various components of UIMA. It is also being widely used for information extraction, where the text analysis components of UIMA extracts the information and then the data is analyzed using various business intelligence tools [852]. Currently UIMA is used by IBM Watson to analyze unstructured data and IBM also uses it for its search platform. cTAKES uses it to describe

its patient physician encounters by analyzing the narrative text. DKPro Core is another NLP open source software which is based on Apache UIMA [395].



title	VirtualBox
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

VirtualBox ??? was developed by Innotek GmbH and is now run by Oracle Corporation. This technology is a hosted hypervisor, meaning that it can create and run virtual machines. VirtualBox acts as the host machine and is designed for x86 computers[853]. It may be installed on OpenSolaris, Windows, Linux, and macOS. This means that you can download VirtualBox on your own computer and then virtually run multiple different operating systems. For example you can download VirtualBox on a macbook and run MacOS as well as AMD and Intel OS through virtual machines. This is a very attractive design for developers who want to build software on their own personal computers [854]. There are two forms of emulated environments that VirtualBox can load: software-based virtualization and hardware assisted virtualization. The software-based virtualization runs in rings 0 and 3 of Intel ring architecture. Ring architecture or protection rings support data from outside device drivers that could impair the functionality of the data or the data itself [855]. Using protection rings, VirtualBox can safely run the host virtual machines and guest virtual machines' codes. This means that companies with older hardwares and softwares can use VirtualBox to run more intricate network systems on up to date virtual software. VirtualBox also has a snapshot feature. This can possibly prevent production system releases going wrong and nothing

“gives a server administrator the chills like a production release going wrong” [856].

The snapshot feature allows a developer to take back the system to

the point of time that the snapshot was taken. This can help with debugging. VirtualBox also enables ease of sharing between virtual machines through a drag and drop feature that runs over the GUI. This can lead to less mistakes with larger sums of data sharing.

3.388 VMware ESXi



title	VMware ESXi
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

VMware ESXi (formerly ESX) is an enterprise-class, type-1 hypervisor developed by VMware for deploying and serving virtual computers [857]. The name ESX originated as an abbreviation of Elastic Sky X. ESXi installs directly onto your physical server enabling it to be partitioned into multiple logical servers referred to as virtual machines. Management of VMware ESXi is done via APIs. This allows for an agent-less approach to hardware monitoring and system management. VMware also provides remote command lines, such as the vSphere Command Line Interface (vCLI) and PowerCLI, to provide command and scripting capabilities in a more controlled manner. These remote command line sets include a variety of commands for configuration, diagnostics and troubleshooting. For low-level diagnostics and the initial configuration, menu-driven and command line interfaces are available on the local console of the server [858].



title	Voldemort
status	10
section	NoSQL
keywords	NoSQL

"Voldemort is a distributed data store that is designed as a key-value store used by LinkedIn for high-scalability storage. It is named after the fictional "Harry Potter villain Lord Voldemort"[859].

It supports a pluggable architecture which allows the support of multiple storage engines in the same framework. This allows us to integrate a fast, fault-tolerant online storage system, with the heavy offline data crunching running on Hadoop.

3.389 Key features:

- Data is automatically replicated over multiple servers .
- Data is automatically partitioned so each server contains only a subset of the total data. This helps to handle any server failure
- Pluggable serialization is supported to allow rich keys and values including lists and tuples with named fields
- Data items are versioned to maximize data integrity in failure scenarios without compromising availability of the system
- Each node is independent of other nodes with no central point of failure or coordination
- Good single node performance: you can expect 10-20k operations per second depending on the

machines, the network, the disk system, and the data replication factor [860]

Voldemort is not a relational database, it does not attempt to satisfy arbitrary relations while satisfying ACID properties. Nor is it an object database that attempts to transparently map object reference graphs. Nor does it introduce a new abstraction such as document-orientation. It is basically just a big, distributed, persistent, fault-tolerant hash table. For applications that can use an O/R mapper like ActiveRecord or Hibernate this will provide horizontal scalability and much higher availability but at great loss of convenience. For large applications under internet-type scalability pressure, a system may likely consist of a number of functionally partitioned services or apis, which may manage storage resources across multiple data centers using storage systems which may themselves be horizontally partitioned. For applications in this space, arbitrary in-database joins are already impossible since all the data is not available in any single database. A typical pattern is to introduce a caching layer which will require hashtable semantics anyway. For these applications Voldemort offers a number of advantages:

Voldemort combines in memory caching with the storage system so that a separate caching tier is not required (instead the storage system itself is just fast). Unlike MySQL replication, both reads and writes scale horizontally Data partitioning is transparent, and allows for cluster expansion without rebalancing all data. Hence there is clear separation of storage and logic.

“Data replication and placement is decided by a simple API to be able to accommodate a wide range of application specific strategies [861].”

The storage layer is completely mockable so development and unit testing can be done against a throw-away in-memory storage system without needing a real cluster (or even a real storage system) for simple testing Only efficient queries are possible, very predictable performance It uses key-value storage and use a dictionary to find information

```
value = store.get(key)store.put(key, value)store.delete(key)
```

3.389.1 Disadvantages of Voldemort:

- No complex query filters
- All joins must be done in code
- No foreign key constraints
- No triggers

3.390 VoltDB



title	VoltDB
status	90
section	In-memory databases/caches
keywords	In-memory databases/caches

VoltDB is an in-memory database. It is an ACID-compliant RDBMS which uses a shared nothing architecture to achieve database parallelism. It includes both enterprise and community editions. VoltDB is a scale-out NewSQL relational database that supports SQL access from within pre-compiled Java stored procedures. VoltDB relies on horizontal partitioning down to the individual hardware thread to scale, k-safety (synchronous replication) to provide high availability, and a combination of continuous snapshots and command logging for durability (crash recovery) [862]. The in-memory, scale-out architecture couples the speed of traditional streaming solutions with the consistency of an operational database. This gives a simplified technology stack that delivers low-latency response times (1ms) and hundreds of thousands of transactions per second. VoltDB allows users to ingest data, analyze data, and act on data in milliseconds, allowing users to create per-person, real-time experiences [862].

3.391 vSphere and vCloud fa18-523-85

title	vSphere and vCloud
status	10
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

vSphere was developed by VMware and is a cloud computing virtualization platform. [863] vSphere is not one piece of software but a suite of tools that contains software such as vCenter, ESXi, vSphere client and a number of other technologies. ESXi server is a type 1 hypervisor on a physical machine of which all virtual machines are installed. The vSphere client then allows administrators to connect to the ESXi and manage the virtual machines. The vCenter server is a virtual machine that is also installed on the ESXi server which is used in environments when multiple ESXi servers exist. Similarly, vCloud is also a suite of applications but for establishing an infrastructure for a private cloud. [864] The suite includes the vsphere suite, but also contains site recovery management for disaster recovery, site networking and security. Additionally, a management suite that can give a visual of the infrastructure to determine where potential issues might arise.

3.392 Whirr



title	Whirr
status	90
section	Interoperability
keywords	Interoperability

Apache Whirr is a set of libraries for running cloud services, which provides a cloud-neutral way to run services [865]. This is achieved by using cloud-neutral provisioning and storage libraries such as jcclouds and libcloud. Whirr's API should be built on top these libraries and is not exposed to the users. It is also a common service API, in which the details of its working are, particular to the service. Whirr provides smart defaults for services by which any properly configured system can run quickly, while still being able to override settings as needed. Whirr can also be used as a command line tool for deploying clusters. It uses low level API libraries to work with providers which was mentioned in the [866].



title	Winery
status	10
section	DevOps
keywords	DevOps

Eclipse Winery is a web based environment which allows to model Topology and Orchestration Specification for Cloud Applications (TOSCA) topologies and manage these topologies. It has a graphical user interface and allows users to create and modify TOSCA elements. Winery has four parts -

- Type and template management
- Modeler
- TOSCA modeler
- The repository

The Type and template management allows to manage all TOSCA elements, their types and templates like relationships, node types and virtual machine artifacts. Next is the topology modeler which allows to create service templates, which are nothing but instances of node types and node relations. The Modeler is a web-based application to create BPMN models. It can support BPMN4TOSCA. The repository is a system to store and manage TOSCA models and it can also facilitate importing and exporting tasks [867]. Some of the functionalities that are provided by Winery tool are -

- Consistency checker to check if a service template is valid.
- XaaS Packager for deploying web application by reusing the existing templates.
- Topology Completion allows users to model an incomplete service template or model.
- Splitting and matching allows to split and match function

facilities for rearrangement of application components, Key based policy template generator allows to generate security policy template,

- Implementation Artifact Generator allows to specify the function of a node type
- Compliance Checking allows the topology compliance checking of Winery to tell about the constraints and requirements for topology templates, which is in reusable form of topology based compliance rules.

Winery structure can be described as Databases consisting of Types, Templates and Artifacts. The repository REST interface consists of repository, TOSCA Importer and TOSCA Exporter. The graphical user interface or GUI portion consists of Type, Template and Artifact management GUI, Topology Modeler GUI and Plan Modeler GUI components. Eclipse winery allows us to create service templates as directed graphs using TOSCA model editor. The service templates help us annotate requirements, Artifacts, properties and policies. Modeled service templates can be exported based on XML standard because the Winery data model is based on XML standard. These services helps us in enabling the importing and exporting processes using the TOSCA XML Transformer Model Importer and Exporter. The Cloud Service Archive (CSAR) package in the backend hosts all these service components. It is also used to deploy cloud applications. The BPMN4TOSCA Management Plan Editor helps us create or modify BPMN models on a web based user interface. It also helps us to load the existing management plans to Winery. Thus, Winery on a whole is a comprehensive package having services, management systems and user interface to handle model topologies and cloud applications [868].



title	Wink
status	90
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Wink is a Java based framework which allows a user to develop and work with RESTful web services [869]. This tool was built by implementing the JAX-RS framework; the Java API which provides support for creating web services [870]. Wink provides both a client and server module. The server module facilitates the development of REST services and the client module consumes these services [871]. The benefit that Apache Wink provides is that the framework makes it easier for the developer by separating out the low-level details and the business application [871]. This allows the developer to focus on the details and logic of the application being developed rather than spending time sorting through the technical aspects of REST web services.

Wink uses building blocks to contain particular REST components [871]. The service implementation building blocks consist of HTTP methods, URL query parameters, URL handling and URI dispatching to name a few [871]. The client building blocks contain resources such as client request, client response, input and output stream adapters. Lastly, there is a collection of Wink runtime building blocks to deploy the process.

Apache Wink was moved to the Apache Attic in April 2017 [872]. This means that the project has reached its end-of-life and that it is no longer an active project with the Apache Software Foundation. The Apache Attic is designed to be non-impactful to users meaning that applications developed using the existing Wink framework will still work. However, the project will not have any future releases or bug

fixes meaning that an application developer will want to explore other frameworks to develop RESTful web services. Another Apache supported option to develop web services is Apache CXF. This tool supports RESTful services and is actively being updated. Some additional tool for developing RESTful web services are Spring MVC, Restlet and Jersey [873].

3.395 Xcat



title	Xcat
status	10
section	DevOps
keywords	DevOps

xCAT is defined as extreme cloud/cluster administration toolkit. This open source software was developed by IBM and utilized on clusters based on either linux or a version of UNIX called AIX. With this service administrator is enabled with a number of capabilities including parallel system management, provision OS usage on virtual machines, and manage all systems remotely [874]. xCAT works with various cluster types such as high performance computing, horizontal scaling web farms, administrative, and operating systems. [875]

3.396 Xen



title	Xen
status	90
section	IaaS Management from HPC to hypervisors
keywords	IaaS Management from HPC to hypervisors

Xen is the only open-source bare-metal hypervisor based on microkernel design [876]. The hypervisor runs at the highest privilege among all the processes on the host. Its responsibility is to manage CPU and memory and handle interrupts [877]. Virtual machines are deployed in the guest domain called DomU which has no access privilege to hardware. A special virtual machine is deployed in the control domain called Domain 0. It contains hardware drivers and the toolstack to control the VMs and is the first VM to be deployed. Xen supports both Paravirtualization and hardware assisted virtualization. The hypervisor itself has a very small footprint. It is being actively maintained by Linux Foundation under the trademark XEN Project. Some of the features included in the latest releases include {em Reboot-free Live Patching} (to enable application of security patches without rebooting the system) and KCONFIG support (compilation support to create a lighter version for requirements such as embedded systems) [878].

3.397 Yarcdata



title	Yarcdata
status	10
section	NoSQL
keywords	NoSQL

Yarcdata is Cray subsidiary providing Analytics products, namely the Urika Agile Analytics Platform and Graph Engine. Cray's Urika (Universal RDF Integration Knowledge Appliance) system is a hardware platform designed specifically to provide high-speed graph-retrieval for relationship analytics [879]. Urika is a massively parallel, multi-threaded, shared-memory computing device designed to store and retrieve massive graph datasets. The system can import and host massive heterogeneous graphs represented in the resource description framework (RDF) format and can retrieve descriptive graph patterns specified in a SPARQL query.

Urika-GD is a big data appliance for graph analytics helps enterprises gain key insights by discovering relationships in big data [880]. Its highly scalable, real-time graph analytics warehouse supports ad hoc queries, pattern-based searches, inferencing and deduction. The Urika-GD appliance complements an existing data warehouse or Hadoop cluster by offloading graph workloads and interoperating within the existing analytics workflow

Cray Graph Engine is a semantic database using Resource Description Framework (RDF) triples to represent the data, SPARQL as the query language and extensions to support mathematical algorithms [881].

The paper Graph mining meets the semantic web outlines the implementation of graph mining algorithms using SPARQL [882].

3.398 Yarn



fa18-523-81

title	Yarn
status	10
section	Cluster Resource Management
keywords	Cluster Resource Management

“YARN - Yet Another Resource Negotiator was initially named as MapReduce 2 or NextGen MapReduce” [883].

The central thought of YARN is to part up the functionalities of asset administration and occupation planning/observing into independent domains. The main goal is to have a universal Resource Manager (RM) and an Application Master (AM) that is available for each application [884]. YARN is the establishment of the new age of Hadoop and is empowering associations wherever to understand a cutting-edge information engineering. Some portion of the center Hadoop venture, YARN is the engineering focus of Hadoop that permits different information preparing motors, for example, intuitive SQL, ongoing spilling, information science and cluster handling to deal with information put away in a solitary stage, opening a totally new way to deal with investigation. Yarn helps new technologies to tap into the power of Hadoop by enabling them to take advantage of the cheap storage and processing.

YARN’s unique reason for existing was to part up the two noteworthy obligations of the Job Tracker/Task Tracker into independent substances [885] - a worldwide Resource Manager, an Application Master for each application, a Node Manager for each hub-slave, one for each application, a Container running on a Node Manager

The Resource Manager and the Node Manager shaped the new conventional framework for overseeing applications in a dispersed way. The Resource Manager is a definitive specialist that parleys assets among all applications in the framework. The Application Manager is very particular about structure, it aids the Resource Manager and Node Manager in tasks like arranges assets, segmenting and executing tasks. The Resource Manager has a scheduler, or, in other words dispensing assets to the different applications running in the bunch, as per limitations, for example, line limits and client limits. The scheduler plans dependent on the asset necessities of every application.

3.399 Zeppelin



title	Zeppelin
status	10
section	Technologies To Be Integrated
keywords	Technologies To Be Integrated

Apache Zeppelin [886] provides an interactive environment for big data analytics on applications using distributed data processing systems like Hadoop and Spark. It supports various tasks like data ingestion, data discovery, data visualization, data analytics and collaboration. Apache Zeppelin provides built-in Apache Spark integration and is compatible with many languages/data-processing backends like Python, R, SQL, Cassandra and JDBC. It also supports adding new language backend. Zeppelin also lets users to collaborate by sharing their Notebooks, Paragraph and has option to broadcast any changes in realtime.



title	ZeroMQ
status	10
section	Inter process communication Collectives
keywords	Inter process communication Collectives

ZeroMQ is an open source library used to build middleware communication systems for applications and software, that require very fast and asynchronous data flow. Initially, it was mainly used for instant messaging and stock trading like real time applications where data flow speed is very important, but later on, the library was developed to support distributed systems, complex networks in communication, transportation, etc. ZeroMQ has APIs to support various high-level languages and it works on most operating systems [887]. The library works through sockets, which are created by following some predefined network communication patterns, which work asynchronously. The suffix of ZeroMQ, MQ suggests that it performs thread queues before processing further. The work of ZeroMQ varies with sockets. Request/Reply pattern used for sending and replying to the messages which were sent. Publish/Subscribe pattern is used distribution of data from publisher or source to multiple recipients, also called subscribers. The Pipeline pattern is for distributing the data to various client nodes and is used to create data pipelines. The Exclusive pair pattern is used for pairing two peers for communication [888]. ZeroMQ also has various transport types like In Process, used for Local In process communication transport, Inter Process used for Local In communication transport, TCP for Unicast communication transport, and PGM for Multi cast communication transportation services [889]. ZeroMQ is different from various other traditional communication packages or tools available outside in the way it can have dual link or the server or client can link, which makes it to be able to wait stably. The client server structure is used in cases where there can be any

communication issues, which are rectified by the ZeroMQ networking utilities. It consists of -

- Streamer for pipelined parallel communication networks
- Forwarder for pub/sub communication connections
- Queue for request and reply services in communication network.

All these features and utilities of ZeroMQ make it a comprehensive library, the services of which are extensively used by large organizations like NASA, AT&T [890].

3.401 ZHT



title	ZHT
status	10
section	NoSQL
keywords	NoSQL

ZHT is a

"a zero-hop distributed hash table" [891].

Distributed hash tables effectively break a hash table up and assign different nodes responsibility for managing different pieces of the larger hash table [892]. To retrieve a value in a distributed hash table, one needs to find the node that is responsible for the managing the key value pair of interest [892]. In general, every node that is a part of the distributed hash table has a reference to the closest two nodes in the node list [892]. In a ZHT, however, every node contains information concerning the location of every other node [893]. Through this approach, ZHT aims to provide

" high availability, good fault tolerance, high throughput, and low latencies, at extreme scales of millions of nodes" [893].

Some of the defining characteristics of ZHT are that it is light-weight, allows nodes to join and leave dynamically, and utilizes replication to obtain fault tolerance among others [893].

3.402 Zookeeper



title	Zookeeper
status	10
section	Monitoring
keywords	Monitoring

Zookeeper provides coordination services to distributed applications. It includes synchronization, configuration management and naming services among others. The interfaces are available in Java and C [894]. The services themselves can be distributed across multiple Zookeeper servers to avoid single point of failure. If the leader fails to answer, the clients can fall-back to other nodes. The state of the cluster is maintained in an in-memory image along with a persistent storage file called znode by each server. The cluster namespace is maintained in a hierarchical order. The changes to the data are totally ordered [895] by stamping each update with a number. Clients can also set a watch on a znode to be notified of any change [896]. The performance of the ZooKeeper is optimum for read-dominant workloads. It is maintained by Apache and is open-source.

4 Incomming

4.1 AlibabaCloud



title	AlibabaCloud
status	95
section	TBD
keywords	TBD

Alibaba Cloud is a tech giant which provides cloud computing services to support both international customers and their own internal business partners who are using Alibaba Group's e-commerce ecosystem.

The service provided by Alibaba Cloud are efficient as they include high-performance and great computing power in the cloud system. Every service offered by them are available as pay-as-you-go along with Anti-DDoS protection and also includes the luxury of Content Delivery Networkss (CDN). On the other hand, Alibaba Cloud is doing a great impact towards research and development of large database systems and advanced big data technologies. Alibaba Cloud's research and development includes Internet of Things technology, virtual reality, smart homes, networking and also cloud-based mobile-device operating systems [897].

4.2 Alluxio



title	Alluxio
status	95
section	TBD
keywords	TBD

Alluxio is open source project under Apache License 2.0. [898] Applications only has to connect with Alluxio to access data stored in any underlying storage systems. Alluxio is Hadoop compatible. In the big data ecosystem, Alluxio [898] lies between computation frameworks or jobs, such as Apache Spark and various kinds of storage systems, such as Amazon S3. It provides fault-tolerance and effective data management across different storage systems through the mount feature. It also has a web-UI for browsing file systems. Alluxio [898] connects the gap between big data applications and traditional storage systems, and expands the set of workloads available to utilize the data.

4.3 AWS API Gateway



title	AWS API Gateway
status	95
section	TBD
keywords	TBD

The AWS API Gateway [899] is used to manage multiple RESTful services in a defined way. You can set up the API Gateway using the CLI, CloudFormation or even Swagger templates. The API Gateway is serverless and AWS will manage all of the underlying infrastructure for you. The design allows you to configure the API mapping and integrations. The API Gateway can then help you define authentication/authorization controls, define the lifecycle for the services and even track transactions for uses like billing.

4.4 Amazon Aurora



title	Amazon Aurora
status	95
section	TBD
keywords	TBD

Amazon's Aurora is a relational database that is compatible with MySQL and PostgreSQL that puts together performance and availability of databases with the power,

"simplicity and cost-effectiveness of open source databases" [900].

Compared to standard MySQL databases, Aurora provides speeds that are up to five times that higher. Its performance is increased by utilizing an SSD-based storage that helps reduce delays and workloads to the system. It is also designed to reduce Input/Output operations and costs so that resources can be available.

4.5 Amazon CloudFront



title	Amazon CloudFront
status	95
section	TBD
keywords	TBD

A Content Delivery Network is a globally distributed network of webservers [901] over the internet at different geographical locations. They form a huge part of internet services today and are deployed at different locations to ensure faster content load times, and lower bandwidths over a network. This technology is highly

“useful to companies that require higher response times” [901]

and distribution of large files to many users at a given time.

It helps accelerate delivery by moving the content close to the end-user therefore reducing hops through the internet. This is often done through caching the content inside a server that is closer to the user. With this, network performance is accelerated, including global presence, and smart computing. Amazon CloudFront is a Content Delivery Network (CDN) that is integrated in Amazons AWS service. It is one of the largest in the world, including others such as Akamai, MaxCDN, and Rackspace. CloudFront is continuing to grow globally and currently

“includes 44 availability zones in 16 different geographic regions today”[902].

This also includes plans for constructing 14 other zones in the coming future.

4.6 AWS CodeStar



title	AWS CodeStar
status	95
section	TBD
keywords	TBD

AWS CodeStar is a developer tool used to develop projects and easily deploy on AWS cloud. It includes all of the tools and services needed for a project development. It supports various templates to set up projects using AWS Lambda, Amazon EC2, or AWS Elastic Beanstalk and IDE platforms such as AWS Cloud9 Eclipse, Visual Studio, CLT. It comes pre-configured with a project management dashboard, an automated continuous delivery pipeline.

Additionally, AWS CodeStar integrates with Atlassian JIRA Software to provide project management and issue tracking system for software project team directly from the AWS CodeStar console [903].

4.7 AWS DeepLens



title	AWS DeepLens
status	95
section	TBD
keywords	TBD

ERROR: CITATION PLACEMENT WRONG WE CAN NOT FIGURE OUT IF THIS MEANS IT IS QUOTED

AWS DeepLens is the world's first wireless high definition video camera which is optimized for Deep Learning. It comes with computer vision model that can be used with the camera. Deeplens integrates with amazon SageMaker and AWS Lambda.

Apart from configuring and running deep learning models, AWS Greengrass can be programmed to run various lambda functions. There are many pre-built models that can run instantly with Deeplens. [904]

Several Features of AWS DeepLens are listed next

- Integrated with AWS
- Build custom models with Amazon SageMaker
- Broad framework support
- Fully programmable
- Custom built for deep learning

4.8 Amazon DynamoDB



title	Amazon DynamoDB
status	95
section	TBD
keywords	TBD

NoSQL refers to a non-relational database that provides high performance and uses various data models such as document, key-value, graph, and columnar. Compared to Non-relational databases, they do not often enforce the use of a schema. DynamoDB is a type of NoSQL database provided by Amazon. It is a cloud-based database that is fully managed and capable of supporting both key-value and document-based models. It comes [905] with a very flexible model and throughput capacity that makes it great for various devices and applications such as those suitable for gaming, IoT, etc. With DynamoDB [906], customers don't have to worry about the burdens of operating distributed services such as hardware setup, configurations, and software patches.

4.9 Amazon EC2



title	Amazon EC2
status	95
section	TBD
keywords	TBD

Amazon Elastic Compute Cloud (Amazon EC2) [907] is a web service provided by Amazon.com. It is a system that allows users to rent cloud computers to run the required applications. It can provide secure service in the cloud, users will be able to run any software or application they want on this virtual machine.

Amazon EC2 has lots of benefits. First, it is inexpensive, it only costs a very low rate for the compute capacity. Second, it is easy to start. It contains several ways to get started with Amazon EC2. Third, Amazon EC2 provides a highly reliable environment.

"The service runs within Amazon's proven network infrastructure and data centers." [907].

What's more, Amazon EC2 and Amazon VPC works together to provide high security.

"Cloud security at AWS is the highest priority." [907].

4.10 Amazon Elastic Beanstalk



title	Amazon Elastic Beanstalk
status	95
section	TBD
keywords	TBD

AWS Elastic Beanstalk [908] is a managed service used for application deployment and management. Using EBS it is easy to quickly deploy and manage applications in the AWS Cloud. Developers simply upload their application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring [909].

Elastic Beanstalk supports applications which are developed in Java, PHP, .NET, Node.js, Python, and Ruby as well as different container types for each language. A container is used to define the infrastructure and technology stack to be used for a given environment [909]. AWS Elastic Beanstalk runs on the Amazon Linux AMI and the Windows Server 2012 R2 AMI provided by Amazon. Initially, it takes some time to create AWS resources required to run the application. User then can have multiple versions of their applications running at the same time. Hence, user can create different environments such as staging and production where each environment runs with its own configurations and resources. AWS Elastic Beanstalk does not have any extra charges. Users need to pay for the resources they have used to store and run the applications such as EC2, S3, RDS or any other resources used.

4.11 Amazon Fargate



title	Amazon Fargate
status	100
section	TBD
keywords	TBD

ERROR: CITATION PLACEMENT WRONG WE CAN NOT FIGURE OUT IF THIS MEANS IT IS PROPERLY QUOTED

AWS Fargate is a technology built on top of Amazon elastic container services and Kubernetes services. It provides container management where there is no requirement of cluster or infrastructure management. Everything is handled at the container level and it scales seamlessly.

Running a container locally over docker is easy but there is a huge overhead in running multiple containers in production like high availability, resiliency, latency, scheduling and resource management. ECS made scheduling and orchestration easy but there are many tasks that it didn't handle like task definition,resource constraints,networking and security. Fargate takes care of most of these tasks except resource definition. Hence all the underlying logistics are taken care by it.

Fargate uses the same task definition schema as ECS and can be launched by ECS APIs [910].

Key features of Amazon Fargate:

- Orchestration
- Enable running containers without server and cluster management

- Eliminates the need to choose server types
- Eliminates infrastructure management
- Seamless scaling

4.12 Amazon Glacier



title	Amazon Glacier
status	95
section	TBD
keywords	TBD

Amazon Glacier is an online file storage web service provided by Amazon which can be used for data archiving and backup [911]. Glacier is part of the Amazon Web Services suite designed for long term storage of data that is accessed infrequently. User can store virtually any kind of data in any format.

Amazon also provides Simple Storage Service for storing and retrieving data but Glacier is much cheaper than S3. As per AWS documentation,

"For Amazon glacier, storage costs are a consistent \$0.004 per gigabyte per month, which is substantially cheaper than Amazon's own Simple Storage Service" [912].

"Customers can store data to Amazon Glacier with a significant saving as compared to on-premise storage. Amazon Glacier is designed to provide average annual durability of 99.99999999% for an archive" [911].

Data is stored in Amazon Glacier as archives. Archives can be deleted at any point of time and billing will be updated accordingly.

Amazon Glacier provides three options for access to archives, from a few minutes to several hours. The AWS Management console is used for Amazon Glacier set up. User can upload and retrieve data programmatically in later phases.

4.13 Amazon Lightsail



title	Amazon Lightsail
status	95
section	TBD
keywords	TBD

Amazon Lightsail is amazon virtual server. It provides virtual private servers which are pre-configured with storage where applications can be deployed and managed. Lightsail provides compute, storage and networking capabilities to deploy and manage web applications in cloud.

LightSail provides SSD based storage, virtual machine,a static IP and DNS management at a very competing price. It has snapshot which has stored backup of each nodes that are set-up. Static IP address can be created which are fixed and assigned to instances. It provides two OS, Amazon Linux and Ubuntu.

Lightsail is packed with range of operating system and application templates that are installed with the creation of an instance. Some of the application templates are Nginx,Node.js, Mean and Lamp.

Each instance of Lightsail gets a private IP address and a public IP address which are accessible over web. [913]

4.14 Amazon Machine Learning



title	Amazon Machine Learning
status	95
section	TBD
keywords	TBD

AmazonML is a service for machine learning that makes it possible to design and build applications that can be used for prediction, forecasting et cetera. It employs algorithms that can help create models that can be used to train and discern patterns from data. A trained model can be used to determine trends that can be used to make predictions when given new sets of data. The Amazon's Machine Learning API

"provides data and model visualization tools, as well as wizards to guide you through the process of creating machine learning models," [914].

It is a service that is highly scalable and can be used to create large numbers of predictions in real-time. The framework also supports Amazon Machine Images (AMI) that provide the resources for faster development of sophisticated models.

"The AMI's are pre-installed with Apache MXNet, TensorFlow, PyTorch, the Microsoft Cognitive Toolkit (CNTK), Caffe, Theano, Torch, Gluon, and Keras to train sophisticated, custom AI models." [915]

4.15 Amazon RDS



title	Amazon RDS
status	95
section	TBD
keywords	TBD

Amazon RDS [916] stands for Amazon Relational Database Service. Amazon RDS gives access to MySQL, MariaDB, Oracle, SQL Server, or PostgreSQL database. It is a managed service provided by AWS which can be used to manage different database administrative tasks. User can select the type of RDS instance and accordingly AWS provides capacity. RDS has capacity to resize as per requirement which enables user to change from one instance type to another instance type without losing its data. It is cost effective and the costing depends on the instance type [917].

“Amazon RDS can automatically backup database and keep that database software up to date with its latest version. RDS makes it easy to use replication to enhance database availability, improve data durability, or scale beyond the capacity constraints of a single database instance for read-heavy database workloads” [917].

High availability is achieved by built-in automated failover from primary database to a replicated secondary database in case of any failure. This replicated secondary database is sync with primary database.

4.16 Amazon Redshift



title	Amazon Redshift
status	95
section	TBD
keywords	TBD

Amazon Redshift is a product of amazon mainly designed as datawarehouse service center that fully manages data warehousing and makes it a very simple and cost-effective application that could be used to analyze all the data using standard SQL interface and existing Business Intelligence tools. Their website gives more details saying

“It allows you to run complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution” [918].

It is an Internet hosting service and data warehouse product that was announced in 2012 and it became a part of the bigger cloud services project namely Amazon Web Services. Few stats about the technology and its adoption -

“It is built on top of technology from the massive parallel processing (MPP) data-warehouse company ParAccel (later acquired by Actian), to handle large scale data sets and database migrations. Redshift differs from Amazon’s other hosted database offering, Amazon RDS, in its ability to handle analytics workloads on big data data sets stored by a column-oriented DBMS principle” [919].

4.17 Amazon S3



title	Amazon S3
status	95
section	TBD
keywords	TBD

Amazon S3 is a simple storage service which mainly focuses on a highly-scalable, reliable, and low-latency data storage infrastructure at low costs [920]. Simple storage Service is a web service provided by Amazon that can be used to store and retrieve data. It can also be used for static website hosting for different web applications. Important feature of S3 is that it is available at any point of time and can be used to store virtually any kind of data in any format [920].

One of the important features of using S3 is that it offers a highly durable, scalable, and secure destination for backing up and archiving critical data [921]. As per AWS documentation,

"Amazon S3 is designed to deliver 99.99999999% durability, and it is used to store data for millions of applications used by market leaders in every industry" [921].

Amazon S3 provides versioning capability to provide even further protection for stored data. It is easy to define lifecycle rules to automatically migrate less frequently accessed data. User can store any number of objects. Total volume is unlimited but one object size can range from 0 bytes to 5 terabytes. With Amazon S3, user needs to pay only for what the usage is. But price vary as per the chosen region of S3.

4.18 Amazon VPC



title	Amazon VPC
status	95
section	TBD
keywords	TBD

A VPC is part of the AWS infrastructure that is logically isolated and spans a whole entire region of AWS to which the instances were created.

"It is one of the most used and famous services inside the Amazon Web Services suite" [922].

It provides the possibility of assigning specific IP addresses, subnets, and network rules when establishing communication with other resources within and out of the infrastructure. This way, the traffic between instances of the elastic cloud compute is protected against network intrusion. This service makes it possible to separate public networks from private providing business enterprises the power to decide which instances can be exposed to the outside word and which cannot. VPC's include six major components that are part of the default VPC. These include gateways, Route Tables, Security Groups, Network Access Lists, and the Classless Inter-Domain Routing (CIDR) Blocks. These resources bring out the simplicity, advanced security, and all the scalability and reliability of AWS [922].

4.19 Ansible



title	Ansible
status	95
section	TBD
keywords	TBD

Ansible is a widely popular open-source tool used for automation of configuration management, application deployment. Ansible is popular because of its simplicity. Originally, Ansible Inc was setup to manage the product. Later in 2015, RedHat acquired Ansible.

“It uses no agents and no additional custom security infrastructure, so it’s easy to deploy and most importantly, it uses a very simple language (YAML, in the form of Ansible Playbooks) that allow you to describe your automation jobs in a way that approaches plain English” [923].

An user doesn’t have to learn a cryptic language to use it. As no agents are required to be installed in the nodes, the tool eases the network overhead. Ansible may use two kinds of server for operation. One is the controlling server that has Ansible installed. The controlling server deploys modules in the nodes through SSH channel. The basic component of Ansible architecture are:

Modules:

This is the unit of work/task in Ansible. It can be written in any standard programming language

Inventory:

Inventory is basically the nodes used

Playbooks:

A play book in Ansible describes in simple language the infrastructure used for the deployment of the tool. This is written in YAML.

4.20 Apache Accumulo



title	Apache Accumulo
status	95
section	TBD
keywords	TBD

Based on Google's BigTable design, Apache has their own data store called Accumulo [924]. Accumulo overlays the Hadoop Distributed File System (HDFS) and Apache Zookeeper. Originally created by the US National Security Agency, Accumulo has a large focus on security and access control. Every key-value pair in Accumulo has its own user restrictions. Accumulo is used mostly in other open source projects and in other Apache projects such as Fluo, Gora, Hive, Pig, and Rya.

Accumulo is a distributed storage system for data, which is simpler than a typical key-value pair system. Each record in Accumulo has the following properties: Key, Value, Row ID, Column, Timestamp, Family, Qualifier, and Visibility. The records are stored across many machines, with Accumulo keeping track of the properties. A monitor is also provided for information on the current states of the system. A garbage collector, tablet server (table partition manager), and tracer (for timing) are also included as well as iterators for data management.

4.21 Apache Ambari



title	Apache Ambari
status	95
section	TBD
keywords	TBD

Ambari is a software to manage Hadoop environment efficiently by providing services like managing, monitoring and provisioning to the hadoop clusters [925].

When Apache Hadoop started developing with aim of increasing its scalability, several application layers started to cover its architecture like Pig, Hive, HBase etc. making the management of Hadoop architecture bulky and unmanageable, and several problems were faced by the developers in handling large hadoop clusters. Ambari is developed aiming to be the solution to the above problems.

Apache architecture includes two main components - Ambari Server and Ambari Agent. Ambari supports 64 bit OS like RHEL 5 (Redhat enterprise Linux), RHEL 6, CentOS 5, CentOS 6 etc [926]. Ambari provided monitoring services through tools like Dashboard views - which shows cluster health and cluster status, also by collecting different metrics like Job status, Maps slots utilization, garbage collection.

4.22 Apache Atlas



title	Apache Atlas
status	95
section	TBD
keywords	TBD

Apache atlas is the novel adaptable platform which incorporates the center set of the functional administration services. The Apache atlas empowers the ventures to effectively meet the prerequisites inside the Hadoop. Additionally, it delivers the integration of the entire data environment. The database researchers, data analysts, and the data administration group can take advantage of the open metadata management and the administration capabilities can be utilized for the organizations to create and make the catalog of their information resources. These resources can be classified and collaborated inside the venture effortlessly [927].

There are three main core components of the Apache Atlas, Type System, Graph Engine and Ingest/Export. The type system enables the modeling of the metadata for the objects that are intended to be administered. The metadata objects are represented by the entities which are the instances of the Types. Inside the Apache Atlas, the metadata objects are managed with the help of the graph model. The rich relationships between the metadata objects are taken care by this approach by providing the good adaptability and effective handling of the relationships. Additionally, the graph engine also provides the effective indexing by creating the relevant indices for the metadata objects with the goal of providing the efficient search results. The next component called ingest helps the users to post the metadata to the Atlas. In contrast, the export component will help the users to expose the metadata of the Atlas and creates an event specific to each change. The end users will be able to respond to these alterations in the real time by consuming these change events [928].

4.23 Apache Avro



title	Apache Avro
status	95
section	TBD
keywords	TBD

Avro [929] is a framework for data serialization, where serialization is a process of translating object or data structure into a format that can be stored. Apache Avro can translate very high datastructure formats. It provides binary data format which is very fast and compact. Avro can also provide Remote procedure calls.

Avro is completely based on schemas. Data is stored in file along with the schema and can be read by any program, since schema is available when ever data is read or written it can be used as dynamic scripting languages.

Avro differs from similar systems like Thrift, Protocol Buffers by schema evaluation, untagged data and dynamic typing.

4.24 Apache Chukwa



title	Apache Chukwa
status	95
section	TBD
keywords	TBD

Chukwa [930] is a data collection system built on top of Hadoop to monitor large distributed file systems. It collects data from various data providers and analyses them using MapReduce. Chukwa inherits Hadoop's scalability and robustness. Chukwa has mainly four components: Relies on data agents. Collectors collect data and gives it to stable storage. This data is parsed and archived using MapReduce jobs. It provides interface to analyse and display results [931].

4.25 Apache CloudStack



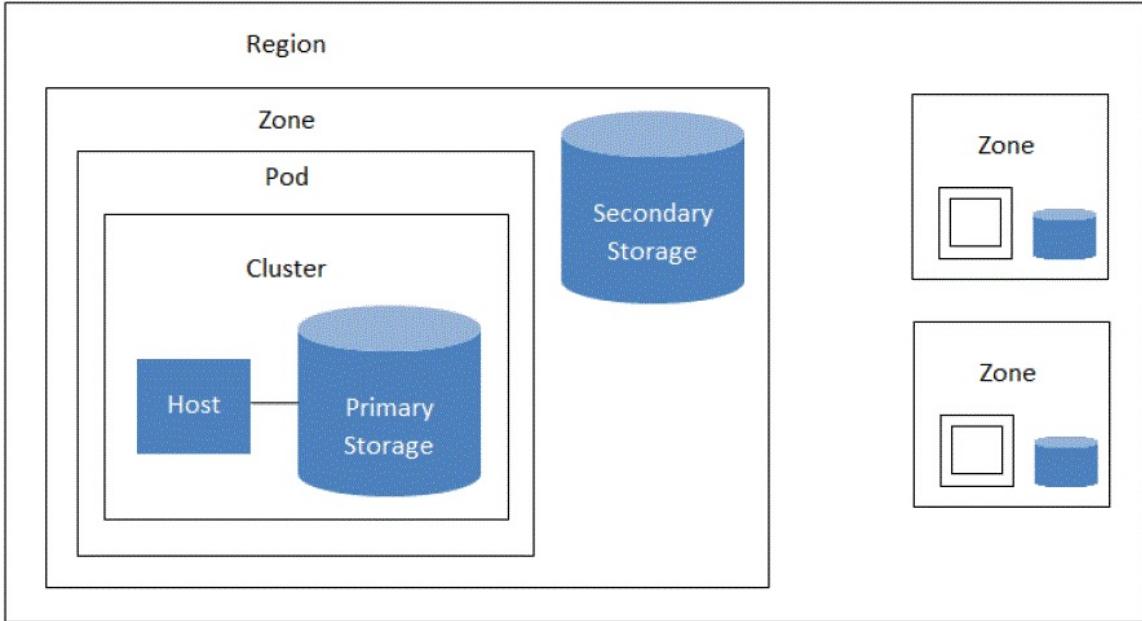
title	Apache CloudStack
status	95
section	TBD
keywords	TBD

Apache CloudStack is an open source that provides a highly scalable and available cloud management platform for IT Enterprises and service providers. CloudStack was originally developed by Cloud.com and was known by the name VMOps. In 2011, Citrix acquired the product and donated it to Apache.

"CloudStack is being developed to help managed service providers and enterprise IT departments create and operate public cloud, private cloud or hybrid clouds with capabilities equivalent to Amazon's Elastic Compute Cloud (Amazon EC2). It uses existing hypervisors such as KVM, VMware ESXi|VMware vcenter and XenServer/XCP for virtualization. In addition to its own API, CloudStack also supports the Amazon Web Services (AWS) API[3] and the Open Cloud Computing Interface from the Open Grid Forum." [932].

The key feature of the product are (1) high availability of resources (2) network management (3) provides GUI for ease of management (4) compatible with most of the hypervisor/virtual monitor (5) it provides the snapshot management. e.g. This feature is very useful is saving a state [snapshot] of a virtual machine. The vm can later be reverted to the stored state. The basic deployment of CloudStack just needs two machines: A server and a hypervisor that is a monitoring system. The process can be over simplified by configuring one machine to serve both the purpose. The same simple system can easily be scaled to a

zone or a pod. Figure [\[F:cloudstack-scalability\]](#) depicts how the simplest deployment infrastructure can be scaled to provide an advanced support system.



CloudStack Scalability [933]

4.26 Apache Couch DB



title	Apache Couch DB
status	95
section	TBD
keywords	TBD

Apache Couch DB is a NoSQL database which uses document instead of tables to store the data. It simplifies the interaction with application as data can be fetched or stored in form of JSON objects [934]. The main advantage of using CouchDB is that it is compatible with variety of application. It can be used to integrate data from web-based applications, mobile applications, web browsers to distributed server clusters. It makes transfer between all these components happen smoothly while providing high performance and totally reliable framework. It also supports Map-reduce operations [934].

4.27 Apache Curator



title	Apache Curator
status	95
section	TBD
keywords	TBD

Apache Curator framework initially developed by NetFlix. In July 2007 this framework has been open sourced to GitHub.

Apache Curator is a collection of Java based client libraries for Apache Zookeeper, a centralized distributed service [935]

It includes a high level API framework and utilities to make using Apache Zookeeper much easier and more reliable. It also includes recipes for common use cases and extensions such as service discovery and a Java 8 asynchronous DSL [935].

The Curator framework consists of set of API's that prominently streamline using Zookeeper. This framework adds various features that build on Zookeeper and handles the complexity of managing connections to the Zookeeper cluster in the distributed environment and retry operations. The benefits of curator framework are: Automatic connectionmanagement, Cleaner API, Recipe implementations [936].

The Apache curator RPC Proxy gives an access to non Java Virtual Machine languages or environments. Organizations can unify the ZooKeeper usage across the environments.

Nirmata workflow is a Apache ZooKeeper service and Apache Curator based library that enables distributed task workflows in the enterprise environments. This service can model simple to complex relationships, manage tasks relationships and distributed job scheduling, it is a simple API, supports run time cluster changes without service

interruption. By default it supports high availability and disaster recovery and no single point of failure in the enterprise [937]

4.28 Apache Delta Cloud



title	Apache Delta Cloud
status	95
section	TBD
keywords	TBD

Apache DeltaCloud was developed in collaboration between Apache Foundation and Redhat to provide a programming application that will facilitate management of different cloud interfaces and It was supporting all the major cloud interfaces.

“Each Infrastructure-as-a-Service cloud existing today[when?] provides its own API. The purpose of Deltacloud is to provide one unified REST-based API that can be used to manage services on any cloud. Each particular cloud is controlled through an adapter called a driver. As of June 2012, drivers exist for the following cloud platforms: Amazon EC2, Fujitsu Global Cloud Platform, GoGrid, OpenNebula, Rackspace, RHEV-M, RimuHosting, Terremark and VMware vCloud” [938].

In 2009, DeltaCloud was developed for the purpose of providing one unified API for the major cloud service.

In 2011, it became a part of the Apache’s top level project.

Unfortunately, in 2015 the project was closed due to inactivity. The application though inactive is chosen for the study to understand the case behind the termination of the project. It is primarily because of lack of popularity RedHat withdrew the sponsorship ultimately resulting in the termination of the project.

4.29 Apache Drill



title	Apache Drill
status	95
section	TBD
keywords	TBD

Apache Drill is an open-source framework for distributing computing on applications handling data-intensive analysis. It is the open-source parallel to Google Dremel for querying very large datasets. Drill is an Apache Top-Level [939] project which enables queries to process on many servers at once over multiple datastores. Drill supports many database systems including MongoDB, Amazon S3, Azure Blob Storage, and Google Cloud Storage [940] and storage file formats including Parquet, JSON, CSV, and TSV in MapR-XD [941].

4.30 Apache Geode



title	Apache Geode
status	95
section	TBD
keywords	TBD

Apache Geode is an in-memory distributed data management platform that provides real-time, consistent access to data-intensive application through extensively distributed cloud architectures and supports high availability and disaster recovery in case of any node failures [942]. Apache Geode initially developed by GemStone Systems and later this framework has been renamed as GemFire.

Gemfire was first installed in the financial sector as the transactional, low-latency data engine used in Wall Street trading platforms [942]. Distributed cache servers are generalization that define the nodes. In each cache we define regions, regions are equivalent to tables in any relational databases or XSD schema structure and manage data in the distributed environment. For high availability the data is replicated to multiple regions (same data is available on each cache servers) by which it ensures high availability as one member goes down still copy is available on other cache member. Locator's responsibility to determine and load balance client (MapReduce, JTA, spring, REST service call, or API) requests to be processed by available cache servers. Locators get notifications continuously if there is any issue in the cluster members, based on this the client request will be navigated appropriately [942]. The main features of this framework are high performance, scalability, fault-tolerance for any data grid platform and can be integrated to other open sources technologies - Spring Data Gemfire [943], Spring Cache [944], and Python [945].

4.31 Apache Ignite



title	Apache Ignite
status	95
section	TBD
keywords	TBD

Apache Ignite is an in-memory distributed database, caching, and processing platform for transactional, analytical, and streaming workloads, delivering in-memory speeds at petabyte scale [946].

Unlike in-memory databases, Apache Ignite works on top of existing databases and requires no rip-and-replace or any changes to an existing RDBMS. Users can keep their existing RDBMSs in place and deploy Apache Ignite as a layer above it. Apache Ignite can even automatically integrate with different RDBMS systems, such as Oracle, MySQL, Postgres, DB2, Microsoft SQL and others. This feature automatically generates the application domain model based on the schema definition of the underlying database and then loads the data. Moreover, IMDBs typically only provide a SQL interface while Apache Ignite provides a much wider ecosystem of supported access and processing paradigms in addition to ANSI SQL. Apache Ignite supports key/value stores, SQL access, MapReduce, HPC/MPP processing, streaming/CEP processing and Hadoop acceleration, all in one well-integrated in-memory data fabric [947].

4.32 Apache Impala



title	Apache Impala
status	95
section	TBD
keywords	TBD

Apache Impala acts as analytic database for Apache Hadoop. Impala can be used from many open source distributions like Cloudera, MapR, Oracle and Amazon. Impala has Enterprise-Class security i.e it is integrated with native hadoop security and Kerberos for authentication [948].

Apache Impala is massively Parallel Processing SQL query engine that works on data stored and run in Apache Hadoop clusters.

“It enables users to issue low- latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation” [949].

4.33 Apache Karaf



title	Apache Karaf
status	95
section	TBD
keywords	TBD

Apache Karaf is a lightweight enterprise application container. This container can be used as a standalone server. It also supports the run anywhere, It requires java to run. It also runs in cloud and Docker images [950].

Apache karaf can be used in any enterprise application development and deployment. It provides similar capabilites of other commercial application servers IBM WebSpere, Oracle Weblogic provides. Capabilities like transactional, log, messages, and web. The disadvantage of using this container as an enterprise server are, there won't be support provided in case of any issues in the environment where as support will be provided for commercial servers like IBM WebSphere and Oracle Weblogic.

As this a light weight container, open source, and provides other features like transaction management [951], logging, security configurations, enterprise application deployment, console access, remote access, and dynamic configuration changes. With this flexibility, Karaf is the perfect solution for microservices, systems integration, big data, and much more [950]. In order to run the Karaf, it requires Java Standard Edition 8 or later. For more details on setup, configuration, deployment and download, please refer index.

4.34 Apache Kylin



title	Apache Kylin
status	95
section	TBD
keywords	TBD

Apache Kylin is an Distributed Analytics Engine which has the capability to query massive SQL data at sub-second speed. It is an Big Data approach to multi-dimensional analysis (OLAP) on Hadoop/Spark. The platfrom has the capability to interface with applications using ODBC/JDBC or Restful API's. The OLAP Cube technology that uses HBase for storage enable the query engine to achieve high speed data querying on tables that contain 10+ billion records.

4.35 Apache Mahout



title	Apache Mahout
status	95
section	TBD
keywords	TBD

Apache Mahout is commercial platform built for scalable implementation of machine learning algorithms. It has support for Apache Spark implementation. The platform built primarily for distributed analytics provides functionalities for clustering, classification, collaborative filtering etc. It provides a linear algebra framework that lets users implement their own algorithm for data analytics. The platform though available commercially is still in its development stage.[952].

4.36 Apache Mesos



title	Apache Mesos
status	95
section	TBD
keywords	TBD

Apache Mesos is the distributed systems kernel built similarly to the Linux kernel, but runs on a different level [953]. Apache Mesos performs container and support and massively scalable data support by splitting scheduling into a two-level architecture. Applications running on Mesos are containerized separately from the framework handling infrastructure scheduling operations [954].

4.37 Apache Phoenix



title	Apache Phoenix
status	95
section	TBD
keywords	TBD

Apache Phoenix [955] is an open-source database engine by Apache that works in tandem with Hadoop and HBase. Because it uses HBase, Phoenix is a noSQL store. The system supports online transaction processing (OLTP). Using SQL and Java database connectivity (JDBC), Phoenix allows for queries over millions of rows to be executed in milliseconds. Phoenix overlays HBase and allows data to be accessed directly via SQL queries (through JDBC). It allows for indexing and parallelization to greatly reduce query time. Phoenix is compatible with the host of Hadoop products such as Spark, Hive, and MapReduce. Phoenix allows table modifications through DDL (Data Definition Language) commands. These commands use simple SQL statements to create or alter tables. Phoenix also uses ACID (Atomicity, Consistency, Isolation, Durability) transactions.

In order to begin using Phoenix, you must first install java and download the Phoenix jar file. Phoenix is currently used by many large corporations such as eBay, Salesforce, and Bloomberg.

4.38 Apache Whirr



title	Apache Whirr
status	95
section	TBD
keywords	TBD

Apache Whirr provides collection of libraries for running cloud services in a neutral way. Whirr began as a set of shell scripts for running Hadoop on Amazon EC2, and later matured to include a Java API based on the Apache jclouds project.

It defines the layout of clusters, It also has scripts to run operations to start, stop and terminate new clusters [956].

4.39 Apache Zookeeper



title	Apache Zookeeper
status	95
section	TBD
keywords	TBD

Zookeeper is a open source centralized service that enables synchronization across cluster. It is also designed to maintain naming, configuration information, and provide group services.

An application can create znode in Zookeeper which can be updated by any node in the cluster and updates on that node can have track of changes to that znode. This kind of znodes are used to keep track of updates in the entire cluster which is how it provides centralized infrastructure [957].

4.40 Apatar



title	Apatar
status	95
section	TBD
keywords	TBD

Apatar [958] is a data integration tool which provides the capability to work with data across different systems and helps to move data between those systems. It also provides ETL capability for the data extraction and transformation. Application point of view it can be used in data warehousing, data migration, synchronization and integration between applications. It can be used across heterogeneous systems like databases, files, FTP, Queue, and applications like ERP, CRM. Since it is an open source tool developed in Java, it provides platform independence and can be used on any operating system. It provides flexible deployment options as desktop, server or embedded into a 3rd party software. The desktop deployment comes with a GUI client installation along with command line support on the local machine. Server deployment allows Apatar to be deployed as server engine over the network. The embedded option allows other software providers to embed Apatar into their software to provide data integration capabilities. Apatar has GUI for mapping and design which can be used by technical as well as the non-technical person. Apatar is based on modular open application architecture which allows customization and flexibility to modify the source code for customized business logic or integration with new systems. As per the Apatar website, it currently supports connectivity and works with Oracle, MS SQL, MySQL, Sybase, DB2, MS Access, PostgreSQL, XML, InstantDB, Paradox, BorlandJDataStore, CSV, MS Excel, Qed, HSQL, Compiere ERP, SalesForce.Com, SugarCRM, Goldmine, any JDBC data sources and more. Apatar also has data quality tool which helps with the data cleansing. It provides support to multiple languages as it is Unicode compliant. The Apatar

architecture consists of 3 major component as presentation/GUI, ETL, and data source. GUI is used to perform various data integration task like data mapping, data source configuration etc in a user-friendly way. Data source provides various connectors to connect with different data sources like databases, files, application (SAP, Siebel, etc), real-time feeds like queuing services. Extract, Transformation and Load (ETL) component provides functionality like data transformation, real-time in-memory data processing, data cleansing and validation, data exception/rejection management, data loading, post data load processing like archival, indexing, aggregation and scheduling and event management.

4.41 AppFog



title	AppFog
status	95
section	TBD
keywords	TBD

AppFog which acts as platform-as-a-Service (PaaS) is developed on the basis of the Cloud Foundry by Century Link. It empowers developers to center on writing advanced cloud-based applications without having to stress around overseeing the basic foundation. The end result is expanded deftness and efficiency, more proficient use of resources and low operational overhead [959].

Rather than investing time on provisioning servers, setting up databases, designing web servers or updating firewalls. AppFog clients essentially convey their cloud-native applications in an extremely quick, tough, multi-environment PaaS. AppFog underpins the most prevalent runtimes and Systems, simplified application scaling, self load-balancing and many other functionalities. Additionally, with the advantage of platform-as-a-Service, it guarantees tremendous use cases to engineers who require less time to release in order to get all the ends fulfilled with today's strict deadline showcase requests [959].

The portability is enabled by the AppFog by providing the compatibility at the core level. The applications that are offered by other cloud foundry providers can be migrated and incorporated in the same environment. The third party services such as the database, notification services and the key value store services can be integrated into the existing application using the Cloud Foundry's User Provided Service Capability [959].

4.42 Appscale



title	Appscale
status	95
section	TBD
keywords	TBD

Appscale developed with the objective of releasing, sending and scaling the Google App Engine applications over the public and private systems of the cloud, provides the scalable open-source cloud computing. In addition, appscale also provides the clusters on the same environment and comprehensive bolster for programming languages such as Go, Java, PHP, and Python applications. This is enabled with the effective modeling of the AppScale with the App Engine APIs [960].

To enable running applications on any cloud infrastructure Appscale provides the API-based development environment and quick responsive functionalities to the designers. The application rationale and the service system are decoupled from each other in order to effectively control the application release, data storage, resource utilization, backup and migration [960].

Appscale provides a simplified serverless platform for the wide variety of the web and mobile applications. The enterprises that use this platform will achieve the goals to quickly manage the time, cut out the functional costs, improve application stability, and the compatibility to combine the existing platform with the other novel technologies [961].

4.43 Apttus



title	Apttus
status	95
section	TBD
keywords	TBD

Apttus provide various products for Customer Relationship Management (CRM) and use artificial intelligence (AI) to maximize the customer revenue. These product help customer to automate the process and maximize the revenue with artificial intelligence (AI) technology. Apttus product include Enterprise Contract Management, Quote-to-Cash, Configure Price Quote, Business to Business (B2B) E-Commerce, Buy-Side Contract Management and Revenue Management. These are the industry standard process and Apttus help to automate these process by using artificial intelligence. Apttus Quote-to-Cash artificial intelligent product used for automating the end to end process of customer's intent to buy the product and ultimately customer buying the product. The Quote-to-cash process start with the buyer's intention to buy the product and ends with the cash in the bank of your organization. Quote-to-Cash artificial intelligent product take care of all the steps involved in the Quote-to-cash process and provide full automation of customer relationship life cycle. These steps requires very less human intervention. This product help in increasing the revenue of the customer as Quote-to-Cash is heart of the business. Apttus provide both on premise and cloud products [962].

4.44 ArangoDB



title	ArangoDB
status	95
section	TBD
keywords	TBD

ArangoDB is a NoSQL database system used to support multiple data models against a single backend engine. ArangoDB supports three main models which are key-value pair, document and graph. Compared to MongoDB, which is a document oriented database, ArangoDB has added benefits such as scalability, lower operation costs, supporting JOINs and complex transactions [963]. ArangoDB uses its own query language AQL, which is similar to SQL, but has the benefit of querying a schema free database [964]. ArangoDB provides flexibility in terms of querying the data because AQL can be used to query across all supported data models. This ease of use in ArangoDB allows developers to represent the components of their systems by models that are much more suitable. This is the reason for the gain in popularity for native multi-model databases such as ArangoDB [965].

4.45 Amazon Athena



title	Amazon Athena
status	95
section	TBD
keywords	TBD

Amazon Athena [966] is a service from AWS that allows the user to analyze their data stored on Amazon S3 using SQL code. It was created with the purpose of allowing anyone with SQL skills to quickly analyze large datasets. Athena will allow a user to run on demand SQL queries without the need to load or gather the data outside of S3 and can process structured, semi-structured and unstructured data sets. It is serverless so there is no need to deal with the setup or managing of infrastructure. This also allows Athena to scale automatically in order to be able to handle large datasets and complex queries. Athena utilizes Presto which is an open source SQL query engine designed to query data wherever it is stored. You can access Athena in multiple ways including the AWS Management Console, API or JDBC driver. It also integrates into Amazon Quicksight allowing you to visualize the data stored in your S3 environment based on your Athena queries. Athena is great for fast on demand querying, but can be used for complex joins, window functions and arrays as well.

4.46 AtomSphere



title	AtomSphere
status	95
section	TBD
keywords	TBD

Boomi AtomSphere is basically an integration platform that supports all application integration processes between cloud platforms, SaaS and local systems as well. Boomi AtomSphere allows its customers to design cloud based processes called Atoms, which defines the necessities for the integration. It can dedicate

“separate environments for testing, perform parallel processing, message based queuing is a part of its service” and it also allows its run time engines to cluster [967].

4.47 Azure



title	Azure
status	95
section	TBD
keywords	TBD

Azure can support different kinds of operating systems and programming language, which is cloud services that developers can use to get their apps to market faster. And Azure can be trusted, which is secure, private and recognized as the most trusted cloud [968]. Compared to AWS, Azure is the better choice all over the world as the most trusted cloud [969].

Besides, lots of choices can pick in the cloud with Azure.

“Get support for infrastructure as a service (IaaS) on Linux, Java, and PHP Web application platforms. Develop and test your Linux and open source components in Azure. You bring the tools you love and skills you already have, and run virtually any application, using your data source, with your operating system, on your device [970]”.

4.48 Azure Blob Storage



title	Azure Blob Storage
status	95
section	TBD
keywords	TBD

Microsoft Azure BLOB storage service can be used to store and retrieve Binary Large Objects (BLOBs), or what are more commonly known as files [971]

This service can be used to share files with clients and to off-load some of the static content from web servers to reduce the load on them. Azure BLOB storage also provides persistent storage. By using this service, developers get dedicated virtual machines to run code without having to worry about managing those virtual machines. Azure BLOB Storage can store any type of file, such as Image files, database files, text files, or virtual hard drive files. However, when they are uploaded to the service they are stored as either a Page BLOB or a Block BLOB depending on how one plans on using the file or the size of the file. Page BLOBS are optimized for random reads and writes so they are most commonly used when storing virtual hard drive files for virtual machines. Each Page BLOB is made up of one or more 512-byte pages of data, up to a total size limit of 1 TB per file. The majority of files would benefit from being stored as Block BLOBS, which are written to the storage account as a series of blocks and then committed into a single file. One can create a large file by breaking it into blocks, which can be uploaded concurrently and then then committed together into a single file in one operation. This allows faster upload times and better throughput. The client storage libraries manage this process by uploading files of less than 64 MB in size in a single operation, and uploading larger files across multiple operations by breaking down the files and running the concurrent uploads. A Block BLOB has a maximum size of 200 GB [971]

4.49 Azure Cosmos DB



title	Azure Cosmos DB
status	95
section	TBD
keywords	TBD

Azure Cosmos DB is a globally-distributed data service that allows elastically scaling throughput and storage across any number of geographical regions while guaranteeing low latency, high availability and consistency [972]. It can support multiple data models using one backend. This means that it can be used for document, key value, relational, and graph models. It is more or less a NoSQL database because it does not rely on any schemas. However, because it uses query language similar to SQL and can easily support ACID transactions, some people have been classifying it as a NewSQL type of database. What differentiates it from other NewSQL databases, however, is that it does not have a relational data model [973]

4.50 Backblaze



title	Backblaze
status	95
section	TBD
keywords	TBD

Backblaze is a cloud backup service [974], providing solutions to business [975] and private users [976]. Backblaze plans do not have a limit for the amount of data you can backup and the software continuously uploads data present on your device.

If the backup is ever needed it can be downloaded. For cases were the amount of data is to much to be downloaded Backblaze offers a service to mail a encripted flash drive or hard drive instead.

4.51 BigML



title	BigML
status	95
section	TBD
keywords	TBD

BigML [977] is a Machine Learning platform focused on delivering a wide range of ML solutions, while aiming to provide a simplified user experience. Considered as a MLaaS, BigML integrates with most Cloud storage systems in order to load data and train machine learning models to develop predictive insights. BigML is also platform-agnostic, and can utilize existing cloud services, such as 'AWS S3, MS Azure, Google Storage, Google Drive, Dropbox, etc' in order to import data [977]. The BigML platform can handle machine learning tasks such as classification and regression, anomoly detction, cluster analysis, association discover, time series, and topic modeling [978].

The platform also includes a large number of visualizations and offers both a Web UI and API interfaces to deliver these machine learning services. The data sets are reusable, and can be used to build a number of models and ensembles to improve predictive performance. It is also possible to convert these models into procedural code that is offered in various different languages and formats [977].

4.52 Apache BigTop



title	Apache BigTop
status	95
section	TBD
keywords	TBD

BigTop [979] is Apache Foundation project for comprehensive packaging, testing and configuration of bigdata components. It supports Hadoop eco system. It packages RPMs and DEBs so that we can manage and maintain Hadoop Cluster. It provides an integrated smoke testing framework. BigTop provides vagrant recipes, raw images and docker recipes to deploy Hadoop from zero.

4.53 Blockchain



title	Blockchain
status	95
section	TBD
keywords	TBD

A blockchain is a continuously growing list of linked records, called blocks. The most recent transactions are recorded and added to it in chronological order, it allows the public to keep track of all transactions in the chain without central recordkeeping. Each computer participating in the blockchain gets a copy of the data which is downloaded automatically.

Blockchains are decentralized by design and provide a new level of the way we trust data. For use as a distributed ledger, a blockchain is typically managed by a peer-to-peer network on the internet collectively agreeing to a protocol for validating new blocks. Once recorded, the data in any given block cannot be changed. One computer trying to tamper a record would have to beat all the computers in the network that are verifying the transactions. Thus tampering is very close to impossibility.

Blockchains

"are an example of a distributed computing system with high Byzantine fault tolerance. Decentralized consensus has therefore been achieved with a blockchain. This makes blockchains potentially suitable for the recording of events, medical records, and other records management activities, such as identity management, transaction processing, documenting provenance, food traceability or voting" [980].

Although, the popular use of blockchain nowadays is with the generation of cryptocurrency, blockchain has more to offer. Big corporations as well the government Institutions have been realizing the potential of the blockchain technology. Recently, the US Department of Health sponsored a contest to the public to propose how can blockchain be used to help improve the health care systems of the country. IBM is looking into how to integrate blockchain in the food industry. Microsoft has partnered with Accenture to use blockchain to store who don't have any form of identification and there are billions of them in the world. Blockchain is here to stay.

4.54 IBM BlueMix



title	IBM BlueMix
status	95
section	TBD
keywords	TBD

BlueMix is a cloud developed by IBM to provide platform as a service (PaaS) to build enterprise level application. In 2017, IBM merged bluemix brand with the IBM cloud brand and now it is known as IBM Cloud instead of IBM Bluexix [981]. All services offered under IBM Bluemix is now available under IBM Cloud and provides over 170+ services. These services are published as infrastructure and platform services. Infrastructure services are consists of Compute, Storage, Network, Security, Containers and VMware. Platform services are consists of Boilerplates, APIs, Application Services, Blockchain, Cloud Foundry Apps, Data and Analytics, DevOps, Finance, Functions, Integrate, IoT, Mobile and Watson. These wide arrays of infrastructure and platform services help create enterprise level of applications. IBM cloud also provides industry-wide solutions in Banking and Finance, Gaming, As Tech, Retail, Healthcare, Telecommunications, Media and Entertainment which can be readily used by the business. IBM Cloud provides pricing options to use its cloud service as free, pay as you go and subscription. It provides various deployment options as on-premises, dedicated private cloud or public cloud [982].

4.55 BMC Multi-Cloud



title	BMC Multi-Cloud
status	95
section	TBD
keywords	TBD

Cloud service introduced a new concept in how to manage IT application and the infrastructure cost and as it matured more, many businesses started adopting cloud solution for their business needs. This resulted sometimes in multiple cloud implementation depending on the business needs. This cloud implementation single or multiple poses challenges in terms of managing their cost, performance, security, automation, visibility, and migration. BMC Multi-Cloud Management solution is specifically built to handle all those challenges and help overall cloud management easy. It provides cost control by providing cloud cost forecast and analysis. Performance monitor provides real-time performance tracking across multiple clouds and provides predictive analytics to keep cloud performance in-check. It provides all assets and dependencies visibility across the clouds which help in inventory and change management. Multi-Cloud security ensures security policy compliance across clouds also it embeds compliance security testing during software development phase. Automation helps automate workload across multiple clouds. It has migration service which helps with the migration to the cloud as well as provides migration plan and simulates forecast annual cost [983].

4.56 Caffe



title	Caffe
status	95
section	TBD
keywords	TBD

Caffe is a deep learning framework developed by Berkeley AI Research. Caffe is optimized for research experiments and industrial applications. Caffe is built in C++ and CUDA with interfaces available in Python and MATLAB [984]. The open source collection of deep learning models is a valuable bundle of tools for research with models including picture pattern recognition, text parsing, and speech composition [985].

4.57 Apache Carbondata



title	Apache Carbondata
status	95
section	TBD
keywords	TBD

As the amount of data we have increases storing and performing analytics of this data becomes increasingly difficult. Apache carbondata is an indexed file format for storing big data that allows faster analysis on a huge amounts of data [986]. Carbondata runs on top of hadoop YARN and spark and can be uses columnar storage, compression and encoding techniques to perform faster queries on the data.

An Apache Carbondata file system consists of groups of data called blocklets and stores information like schema, in the header and footer co-located in HDFS. The Footer is read once to create the index which is later utilized to optimize queries [986].

Apache Carbondata allows operations like creating tables, updating and deleting them and performing queries on these tables [987].

4.58 Cascading



title	Cascading
status	95
section	TBD
keywords	TBD

Cascading is an open source data processing project started in early 2008. Cascading functions as a work flow workhorse within the Apache Hadoop platform and serves as an alternative API to MapReduce. The Cascading Ecosystems includes multiple project extensions for compatibility with multiple languages, platforms, and functions [988]. Originally written in Java, the Cascading platform can be run on any JVM and includes extensions for application development using Domain Specific Languages (DSLs) such as Python, Ruby, Scala or Clojure [989]. Cascading and the Cascading ecosystem were originally designed to be used with the Apache Hadoop MapR distribution for the purpose of developing data-rich applications with analysis and machine learning capabilities [990]. The open source platform and all extensions are available through the Apache Public License.

4.59 CensOS Project



title	CensOS Project
status	95
section	TBD
keywords	TBD

The CensOS Project is a open source project that was developed over the Red Hat Linux system, it is a well maintained open source projects with very low cost for maintainence, which has more than 7 versions updated and maintained. The CenOS Project focuses on developing robust open source ecosystem to personal user, open source community, and companies like Amazon, Google and so on. The CensOS Project provides the both individual users and companies cloud image and powerful cloud developing tools, which enables them to build their own cloud service upon the services offer by the CensOS Project [991].

The CenOS open source project creates the Linux bas distribution cloud system. There are many open source development teams that were grouped by the specific interests. Not only the CensOS itself provides complete documentation of development environment, many development community also provide documentations about their project. Because of the growth of the development community is expanding, the functionality of the CensOS Project will be more advanced in the future.

4.60 Clive



title	Clive
status	95
section	TBD
keywords	TBD

Clive is an open-source, distributed operating system written in Go by the Laboratorio De Sistemas at Universidad Rey Juan Carlos in Madrid [992]. The design goal is to create an environment where applications and services can be compiled along with libraries that permit them to run on bare hardware without a software stack [993]. The design is based on Plan 9, a research system developed at Bell Labs in the late 1980s and first released in 1992[994], and NIX, a

“purely functional package manager” [995] derived from Plan 9 that runs on Linux and Mac OS X.

4.61 Clojure



title	Clojure
status	95
section	TBD
keywords	TBD

Clojure [996] is a fully functional scripting language. Although it is a complied programming language, all its features are available at runtime. Clojure is based on Lisp [997] and uses the same eco system. It provides access to Java framework including hints and type inference. Clojure's main advantage is its implementation of multithreaded programming. Clojure very efficiently breaks a task into subtasks and places them on different JVM threads for parallel processing. Parallel programming has three challenges called three goblins; reference cells, mutual exclusion and dwarves berserkers and Clojure handles them by implementing three tools called futures, delays and promises.

4.62 Cloud AutoML



title	Cloud AutoML
status	95
section	TBD
keywords	TBD

Cloud AutoML is an innovative tool with simple graphical user interface to train and test users custom machine learning models [998]. And these models can be directly used from Google cloud via REST API.

The main purpose of developing Cloud AutoML is to enable users with limited machine learning expertise to train high quality ML models. It is built on Google learning to learn, transfer learning, and Neural Architecture Search technologies.

Google has recently launched first product under Cloud AutoML: AutoML Vision which is a service to access a pre-trained model or create a custom ML models using Cloud ML Engine, for image recognition. It offers drag-and-drop interface to upload images, train and manage models, and then deploy those trained models directly on Google Cloud. For instance, Disney and Zoological Society of London are actively using AutoML Vision [999].

4.63 CloudHub



title	CloudHub
status	95
section	TBD
keywords	TBD

CloudHub is a cloud-based integration platform by MuleSoft which is mainly used for connecting SaaS, cloud and local applications and Application interfaces. CloudHub is an elastic cloud that can scale on demand. We can publish REST API's on it.

"The CloudHub Virtual Private Cloud (VPC) offering enables to construct a secure pipe to on-premise applications through an IPsec VPN tunnel, VPC Peering or Direct Connect" [1000].

CloudHub has a REST API which can perform tasks such as manage, monitor and scale applications.

4.64 Cloudlet



title	Cloudlet
status	95
section	TBD
keywords	TBD

A cloudlet is technique or mechanism by which the cloud capabilities and its wonderful storage,data processing and data analysis power is brought at the edge of the cellular network. The main idea behind cloudlet is to bring the cloud and its services closer to the client (IOT devices,smart phones, smart watches) which cannot independently complete the high load of computation and would require offloadingto meet the computational requirements. This offloading to the main cloud serverwould take relatively longer time in cases where the action has to be taken as soonas possible in real time. Thus in scenarios where latency must be minimum and offloading becomes compulsory we would then be compelled to use cloudlets, where the computation now happens at the edge of cellular network and latency is reduced significantly.

“It is a new architectural element that extends today’s cloud computing infrastructure. It represents the middle tier of a 3-tier hierarchy: mobile device - cloudlet - cloud.” [1001]

Thus a cloudlet can be viewed as a mini data center whose aim is to bring the cloud closer to the Non powerful devices.

“The cloudlet term was first coined by Satyanarayanan and a prototype implementation is developed by Carnegie Mellon University as a research project.The concept of cloudlet is also known as follow me cloud,and mobile micro-cloud” [1001]

4.65 CloudTrail



title	CloudTrail
status	95
section	TBD
keywords	TBD

The AWS CloudTrail [1002] service is an activity recording service provided by Amazon Web Services. The service allows you to track the history of account usage for your AWS instances. The service is not on by default yet when configured, it will record all API calls from all sources like the console, CLI, SDKs or CloudFormation. The data is written into an S3 bucket via JSON and would include attributes like user, IP address, timestamp and the action the user took.

4.66 CloudWatch



title	CloudWatch
status	95
section	TBD
keywords	TBD

The AWS CloudWatch [1003] service is the monitoring service provided by Amazon Web Services. Everything from metrics for resource usage, billing usage, and up to including custom data can be used to group elements into graphs. You can summarize across all instances or you can configure dimensions to allow to focus on certain aspects. Dimensions are a name/value pair that you can establish to target (ex. ServiceName/awskms) yet only certain AWS services are available for aggregation. You can stream the log data to an S3 bucket, to a Lambda function or to Elastic Search. It can also be used to collect logs from your Windows and Linux instances and if you develop an API for your application, it can pull from there as well.

4.67 Microsoft Cognitive Toolkit



title	Microsoft Cognitive Toolkit
status	95
section	TBD
keywords	TBD

The Microsoft Cognitive Toolkit (CNTK) is an open-source project that can be used for implementing distributed deep learning commercially. Per Microsoft,

“CNTK allows the user to easily realize and combine popular model types such as feed-forward DNNs, convolutional neural networks (CNNs) and recurrent neural networks (RNNs/LSTMs)”.

Under the covers, CNTK implements stochastic gradient descent that are automatically parallelized across multiple GPUs and servers [1004]. As of this date, CNTK can be run on both Windows and Linux operating systems.

“CNTK also supports the description of neural networks via C++, Network Definition Language (NDL) and other descriptive languages such as Python and C#” [1005].

4.68 Cognito



title	Cognito
status	95
section	TBD
keywords	TBD

The AWS Cognito [1006] service is used to federate your user registration and their ability to sign into your services. The solution allows you to easily manage user pools and can integrate with multiple SDKs like Java, Python, PHP and Ruby. The client application can be configured to use SAML, OIDC or other backend user directory services. The service is intended to be used in conjunction with AWS IAM and STS.

4.69 ConnectTheDots



title	ConnectTheDots
status	95
section	TBD
keywords	TBD

ConnectTheDots is a Microsoft Open source Technology project which makes it possible to [1007] connect IoT devices and sensors to the Microsoft Azure cloud. It includes a variety of [1008] code samples and guides, including a sample end-to-end weather alert solution that uses an Arduino board, a Raspberry Pi and several Azure services. Operating System can be Windows or Linux. Azure IoT Hub converts the JSON string from Sensors and displays a chart. ConnectTheDots [1008] provides a Multi-protocol Gateway to collect data from devices that cannot, or should not, target the cloud directly

4.70 CouchDB



title	CouchDB
status	95
section	TBD
keywords	TBD

CouchDB[1009] is a database designed for web, which use JSON as the file format to store data. You can use web browser to get access to the documents via HTTP. You can use JavaScript to query, combine, and transform your documents. CouchDB is suitable to work with modern web and mobile apps. CouchDB's incremental replication helps you distribute your data efficiently. You can setup the CouchDB as master-master with automatic conflict detection. CouchDB makes web development a breeze because its suite of features, such as on-the-fly document transformation and real-time change notifications. It even helps use web easily with the administration console, which is served directly out of CouchDB. CouchDB is easy to be distributed scaling, because it's highly available and partition tolerant, but is also eventually consistent. CouchDB puts your data safely with the fault-tolerant storage engine.

4.71 Databricks



title	Databricks
status	95
section	TBD
keywords	TBD

Azure Databricks is founded as an open source project by Microsoft in collaboration with and the creators of Apache Spark and Databricks aiming to help clients with cloud-based big data processing using Apache Spark [1010]. Databricks is closely coupled with Azure to provide easy integration, streamlined workflows, and an interactive workspace which satisfied the requirements of data scientists and data engineers [1011]. Azure Databricks is packaged with the complete open-source Apache Spark cluster technologies such as Spark SQL and DataFrames, Streaming, MLlib, GraphX and Spark Core API [1011]. The main advantage of Azure Databrick platform is that it is a zero-management cloud platform that includes fully managed Spark clusters, an interactive workspace for an exploration and visualization and a platform for powering Spark-based applications [1010]. As Databricks website showcases, Viacom, Shell Energy, HP Inc and Hotels.com are few successful applications which utilizes Databricks services [1010]. Databricks also provides enterprise level Azure security to protect the data using Azure Active Directory integration, role-based controls, SLAs, etc. [1010].

4.72 Datalab



title	Datalab
status	95
section	TBD
keywords	TBD

Cloud Datalab [1012] is an open source tool part of the Google Cloud Platform suite which focuses on delivering analytic, machine learning and visualization solutions. Fully integrated with the Google Platform Suite, Datalab can leverage the data stored in various other Google solutions ranging from BigQuery and Machine Learning Engine to Compute Engine and Cloud Storage [1012]. In addition to this integration, Cloud Datalab delivers a solution for developers to generate reports after exploring, transforming, and visualizing data using Python, SQL, and Javascript. Visualizations are primarily derived from sources such as Google Charting or matplotlib [1013].

Based on Jupyter [1012], Datalab utilizes the existing platform community and supports a ‘large number of existing packages for statistics, machine learning etc’ [1012]. Developers can also extend this open source software by forking it. However, it should be noted that although this is an open source product and is free, cost is a factor once this is deployed as an App Engine application and cloud services are utilized [1013].

4.73 Datameer



title	Datameer
status	95
section	TBD
keywords	TBD

Datameer is self service, schema free Big Data Analytics tool which provides end to end analytics. Datameer, a data analytics application purpose-built for Hadoop. It offers big data integration, analytics, visualization, smart execution, technology, app market, cloud, and smart analytics products. Datameer provides Smart Execution, which selects and combines computation frameworks for Datameer workloads. It also offers Smart Analytics, which provides advanced functions, including clustering, decision trees, column dependencies, and recommendations to find groups and relationships. Datameer is also used in cleansing the data that was ingested or as it is being ingested, and then it has the ability to query the data using Hive/Spark and provide visualization for the queried data. Datameer scales up to thousands of nodes and is available for all major Hadoop distributions [1014]. Datameer specializes in analysis of large volumes of data. Datameer provides end to end solution from data extraction, profiling, cleansing, transforming, merging, securing and finally visualization.

4.74 DBI



	DBI
status	95
section	TBD
keywords	TBD

DBI is a package for R that provides a common interface to databases for R programmers to use [1015]. This allows R to access data that is too big to fit into local memory, or even onto local disk. Key components are classes for database connections, and database results, which can be treated differently, to minimize local computation. Connections to particular database systems, such as MySQL, or PostgreSQL are handled through connectivity packages, such as odbc [1016].

4.75 DBplyr



title	DBplyr
status	95
section	TBD
keywords	TBD

DBPlyr is the bridge between R's immensely popular tidyverse, and the DBI data connection family [1017]. The package allows tables on remote or local databases, regardless of backend, to be treated as first-class data structures in R. It does this by procedurally generating (usually SQL) queries for the databases on the fly [1018]. While the data semantics are agnostic (all data structures are treated the same, regardless of provenance), dbplyr is aware of the limitations of different systems, and will adjust its queries accordingly. Further, dbplyr will evaluate queries lazily, meaning that almost no data is transferred into local memory until it is explicitly asked for.

4.76 Data Virtualization



title	Data Virtualization
status	95
section	TBD
keywords	TBD

Data Virtualization is a modern approach of data integration. Denodo is a data virtualization platform [1019] which provides agile, high performance data integration and data abstraction capabilities. Denodo platform [1019] acts as a logical abstraction between disparate data sources and consuming applications. Denodo allows [1019] intelligent caching for real-time performance. By becoming single virtual layer Denodo Platform [1019] reduces redundancy and resolved quality issues by imposing data model governance. Denodo platform can be used for many use cases covering operational and analytics functions.

4.77 Distributed Machine Learning Tool Kit

title	Distributed Machine Learning Tool Kit
status	95
section	TBD
keywords	TBD

Distributed Machine Learning Tool Kit, or otherwise known as DMTK [1020], is a scalable collection of distributed machine learning algorithms capable of training models on big data sets for increased accuracy. This collection of ML algorithms and computational resources is managed, and has new content added, by Microsoft. The existing DMTK product includes a DMTK Framework for a ‘unified interfaces for data parallelization, hybrid data structure for big model storage, model scheduling for big model training, and automatic pipelining for high training efficiency’ [1020]. DMTK also includes LightLDA (a topic model algorithm), Distribued Word Embedding, and Light GBM (a gradient-boosting tree framework). Developers are also able to utliize the framework for their own custom ML algorithms [1020]. Microsoft has made DMTK open source to encourange ML practitioners and researchers to make contributions to the toolkit [1021].

DMTK’s system innovations regarding its computational resources allow for users to run big data/model algorithms with increased performance. An example of ‘a topic model with one million topics and a 20-million word vocabulary, or a word-embedding model with 1000 dimensions and a 20-million word vocabulary, on a web document collection with 200 billion tokens’ would take a machine cluster of 24 machines as opposed to using thousands [1021]. Capabilities extend beyond just topic modeling and cover various machine learning complex tasks such as speech recognition and computer vision [1021].

4.78 docker



title	docker
status	95
section	TBD
keywords	TBD

Docker is an open platform for developers and sysadmins to build, ship, and run distributed applications, whether on laptops, data center VMs, or the cloud [1022]. It is designed to make it easier to create, deploy, and run applications by using containers. Containers allow a developer to package up an application with all of the parts it needs, such as libraries and other dependencies, and ship it all out as one package. Docker is like a virtual machine but it does not require to create a whole virtual operating system, Docker allows applications to use the same Linux kernel as the system that they're running on and only requires applications be shipped with things not already running on the host computer. This gives a significant performance boost and reduces the size of the application [1023].

4.79 Dokku



title	Dokku
status	95
section	TBD
keywords	TBD

Dokku [1024] is a Platform as a Service (PaaS) that runs on a single server which helps build and manage the lifecycle of applications. It is powered by Docker and can be installed on any hardware. Dokku requires minimum of 1GB memory and Ubuntu 16.04 x64, Ubuntu 14.04 x64, Debian 8.2 x64 or CentOS 7 x64 for the installation. It supports application deployment through git. Technically Dokku is a set of scripts which combined as build pipeline. It takes input as code and generates the running application. It mostly is written in shell script and provides various features as plugins, for example, config, storage etc. Dokku helps in easy code deployment to the cloud so that developers can concentrate more on application logic [1025].

4.80 Drake



title	Drake
status	95
section	TBD
keywords	TBD

Drake is an R package focused on reproducible research and high-performance computing [1026]. It is an R-centric version of Make. The core functionality of Drake is based on the idea that space is cheaper than time. Therefore, it stores local caches of target objects when they are built, along with the commands that were used to build them. From this, it can build a dependency network, and automatically determine which objects are outdated before the next run, and only build the required objects. Because it is R-focused, it has an advantage over make in that it allows for easy plan expansion, rather than make's requirement for explicit commands and targets. Drake also enables higher performance computing, by allowing users to build multiple targets at once, elevating R past its single threaded default.

4.81 Google Dremel



title	Google Dremel
status	100
section	TBD
keywords	TBD

Dremel is an interactive ad hoc query system. It helps user query large datasets. It give results with faster speed as compared to other traditional technologies[1027].

MapReduce framework and technologies built over it such as Pig, Hive suffers from latency issue. User observes time lag between running job and getting results. Dremel overcomes this by using different architecture.

Execution engine of Dremel uses tree algorithm that provides realtime output of queries with high performance.

Dremel is very useful in executing queries over large multi level nested data. Dremel provides very fast SQL like interface to the data. It provides structured query language like syntax. SQL is widely adapted and extensively used by developers for writing queries. This helps developers avoid learning new language for querying.

Dremel run aggregation queries on large datasets with high performance. It achieves this by maintaining hierarchical execution plan. Dremel first arranges execution units in column layout format. It then combines them at multiple levels of trees, thus providing high performance. It is capable of scaling up to multi thousands of CPUs and petabytes of data. [1027].

4.82 Druid



title	Druid
status	95
section	TBD
keywords	TBD

Druid is a high-performance, column-oriented, distributed datastore written in Java. Druid can be used to analyze large volumes of real time streaming data as well as historic data. Druid is horizontally scalable in a cost effective manner and also has the ability to support multi-tenant applications [1028]. In addition to the above mentioned key features Druid also includes the capability to execute fast Online Analytical Processing (OLAP) queries and is fault-tolerant in nature [1029]. The Druid cluster is built from components such as Druid segments, historical nodes, coordinator nodes and broker nodes. Druid also has been identified as a very fast analytics database for fast real-time applications [1030].

4.83 IBM Data Science Experience



title	IBM Data Science Experience
status	95
section	TBD
keywords	TBD

IBM has compiled data science tools in one location, called Data Science Experience. This one location provides access to the IBM Cloud, and allows a customer to run applications from the public or private cloud [1031]. Data Science Experience also allows desktop operations of popular tools[1031]. IBM's Watson computing platform is available in Data Science Experience. Machine learning models through Watson utilizing its vast computing resources. Additionally, open source applications and technologies, such as Python, R, and Apache Spark give users a robust data science toolkit from which to work [1031]. Every aspect of data science is available in Data Science Experience. Visualizations of results are a part of Data Science Experience. Tools such as PixieDust and Brunel are available, without programming experience [1031].

4.84 Edge Computing



title	Edge Computing
status	95
section	TBD
keywords	TBD

Edge computing is a network architecture concept where in the cloud computing capabilities are carried out at the edge of cellular network where the end device or requester is located.

The main idea behind edge computing is to reduce the network latency and radio network resource consumption by bringing the cloud services closer to the device so that latency is reduced significantly.

This mechanism requires leveraging or using resources that may not be connected to a network with devices such as laptops, smartphones, tablets and sensors.

"Edge computing covers a wide range of technologies including wireless sensor networks, mobile data acquisition, mobile signature analysis, cooperative distributed peer-to-peer ad hoc networking and processing also classifiable as local cloud/fog computing and grid/mesh computing, dew computing, mobile edge computing, cloudlet, distributed data storage and retrieval, autonomic self-healing networks, remote cloud services" [1032]

Majority of its application are realized in IOT and other smart connected ecosystem where emergency is the highest priority and data processing is scarce.

Naive example would be a baby crossing a road and an autonomous

vehicle running over the same road, needs to decide as soon as possible to stop motion in order to save baby's life. It cannot send the data to main cloud server and wait for response which would be time consuming and baby's life would be at jeopardy. Hence edge computing would be really useful and saviour for scenarios where offloading to cloud is considered costly.

4.85 Apache Edgent



title	Apache Edgent
status	95
section	TBD
keywords	TBD

The number of connected devices are constantly increasing. Many of the devices that form the internet of things (IoT) are sensors, or devices that are lightweight and do not have a lot of storage space or processing power. Apache Edgent is a programming model that allows development in Java and Android environments and provides a way to perform analytics locally on the edge devices, thus preventing the need to send data back and forth to servers [1033]. Apache Edgent can help reduce the amount of data needed to be stored as analytics can be performed on the data continuously and, only relevant information such as outliers that need to be recorded or data that needs further analysis that require higher computation resources need to be sent to the server [1034].

Apache Edgent can work along with centralized analytics tools thus providing a way to do more thorough analysis on the IoT system. Edgent can communicate with backend systems using common messaging hubs and communication protocols like MQTT, Apache Kafka, IBM Watson IoT platform, and allows custom message hubs as well [1034].

4.86 Elasticsearch



title	Elasticsearch
status	95
section	TBD
keywords	TBD

The central engine to the Elastic product line, Elasticsearch is a distributed, RESTful search engine designed to grow with growing data. Elastic search is a search engine based on Apache's Lucene search library with the first version being released in early 2010 [1035]. Elasticsearch is capable of searching and storing multiple data types, including numeric data, text, geo, and varying levels of structured data using a schema-free JSON format. In addition to the ability to search in real-time, Elasticsearch is capable of analyzing queried results. Elasticsearch's use is compatible with multiple languages such as Curl, Java, Python, C-Sharp, PHP, Perl, JavaScript, and more [1036]. DB-Engines, a

"Knowledge Base of Relational and NoSQL Database Management Systems" ranks Elasticsearch as the top search engine, ahead of both Splunk and Solr [1037].

4.87 ELK Stack



title	ELK Stack
status	95
section	TBD
keywords	TBD

ELK is one of most powerful and scalable BigData solutions in the current market and is indeed doing pretty good. It can solve many challenging problems with respect to indexing, logging, searching, troubleshooting, storage and reporting.

ELK acronyms three open source projects: Elasticsearch, Logstash, and Kibana.

"Elasticsearch is a search and analytics engine. Logstash is a server-side data processing pipeline that ingests data from multiple sources simultaneously, transforms it, and then sends it to a stash like Elasticsearch. Kibana lets users visualize data with charts and graphs in Elasticsearch" [1038].

ELK is one of the most scalable solutions in field of reporting and indexing where Elastic search is an indexing and database kind of service and Logstash works more like a tool for logging everything feeding it to Elastic search for indexing and storing in the database, while Kibana is a nice GUI that helps in data visualization and also allows users to build their own reporting requirements in the Kibana framework which also provides flexibility and scalability. Thus Elasticsearch, logstash and Kibana is a wonderful open source that has collaborated solution for most of problems dealing with BigData and cloud.

4.88 Amazon EMR



title	Amazon EMR
status	95
section	TBD
keywords	TBD

Amazon EMR [1039] is a Hadoop framework that allows the user to process data on the AWS platform using their EC2 technology to spread the load across multiple EC2 instances. Elasticity a major benefit of this product as it can be set to auto scale up or down the number of EC2 instances that EMR is running in a cluster. The user can choose to run a few additional frameworks supported on EMR in addition to Hadoop, such as Spark, HBase, Flink and Presto. It allows the user to focus on the processing of the data and not have to deal with the setup, management or tuning of a Hadoop cluster. Using EMR allows a user to setup and provision a Hadoop cluster quickly and you can scale your compute resources up or down as needed. You can interact with EMR through a web service interface or you can also use the AWS Management Console to launch and monitor your clusters.

4.89 ESRI



title	ESRI
status	95
section	TBD
keywords	TBD

Environmental Systems Research Institute (ESRI) offers geospatial related data services and process online through its proprietary API. Features of the ESRI platform include access to basemaps, geocoding, demographic data, a dynamic world atlas, and multiple data sets in a open-data resource [1040].

4.90 Ethereum



title	Ethereum
status	95
section	TBD
keywords	TBD

Ethereum is an open-source, public, distributed computing that

“runs smart contracts: applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third-party interference” [1041].

A smart contract is a computer protocol intended to digitally facilitate, verify, or enforce the negotiation or performance of a contract without third parties which means there is no need for middlemen like lawyers or notaries. In theory, this means that you can carry out transactions without the waiting times inherent to paper filings, and without paying fees to whoever would typically oversee such a transaction. This is particularly important for people living in countries where the legal system is corrupt, or woefully inefficient. These transactions are trackable and irreversible.

Although smart contract technology is still a very new, it has a wide range of potential applications like as voting, global supply chains, medical records, the financial system. It is very promising it is believed that its full potential has yet to be discovered.

One application of Ethereum is the generation of Cryptocurrency. There are still a lot of doubts regarding the usefulness of cryptocurrency. A lot of people are still on the fence whether they use cryptocurrency as a real currency or not. However, the blockchain technology on which Ethereum is running is a real technology that has a lot of practical applications. Big companies and government

institutions are investing in the blockchain technology and Ethereum can play a major role in it.

4.91 Firebase



title	Firebase
status	95
section	TBD
keywords	TBD

Firebase is an open source project found by James Tamplin and Andrew Lee in 2011 and later acquired by Google in 2014 [1042]. Firebase cloud services started as an online chat message service and soon expanded to provide cloud services such as Firebase cloud messaging, Firebase auth, realtime database, Firebase storage, Firebase hosting, Firebase test lab for Android and iOS and Firebase crash reporting [1043]. A new version of Firebase has released after merging with Google and it provides an unified cloud platform to build Android, iOS, and web Apps [1043]. After the acquisition, Google has stopped supporting their cloud messaging services and merged it with firebase cloud messaging services [1044]. Admob, Analytics, Authentication, Indexing, Test Lab, and Push Notifications are few important features introduced in the latest release of Firebase [1043]. As James [1042] stated, push notification support for Android and iOS mobile application is recently identified as the most famous feature of firebase cloud services.

4.92 Firepad



title	Firepad
status	95
section	TBD
keywords	TBD

Firepad is an open source real-time collaborative code and text editing cloud platform found by Google in 2016 and licensed under MIT [1045]. It is mostly used for rich text editing and code editing as it is empowered with true collaborative editing and intelligent operational transform-based merging and conflict resolution [1046]. Some important features included in the Firepad are cursor position synchronization, undo and redo, text highlighting, user attribution, presence detection and version check-pointing. As Michael Lehenbauer [1045], the founder of Firepad claims that it has no server dependencies and yet provide real-time data synchronization using the Firebase realtime database technology. It is easy to integrate Firepad to any application since inclusion of few JavaScript files would enable the Firepad in all modern browsers such as Chrome, Safari, Opera 11+, IE8+ and Firefox 3.6+ [1045]. As Firepad website showcases, Socrates.io, Nitrous.IO, LiveMinutes, Koding, CoderPad.io and ShiftEdit are few successful applications which utilizes Firepad [1010]

4.93 Fission



title	Fission
status	95
section	TBD
keywords	TBD

Fission [1047] is an open source, serverless framework for Kubernetes. It allows you to create HTTP services on Kubernetes from functions and can help make Kubernetes easier to work with by allowing a user to create services without having much knowledge Kubernetes itself. Fissions method of making things easier for the user is to allow the majority of users to be able to work at the source level. It can abstract away containers from the user. To use it, you create functions using a variety of languages and then add them with a CLI tool. Functions are called when their trigger fires and they only consume CPU and memory while they are running. Idle functions consume no resources with the exception of storage. Some of the suggested uses for Fission are chatbots, webhooks, Rest APIs and Kubernetes events. The only languages supported for it right now are NodeJS, PHP, Go, C# and Python.

4.94 Fluentd



title	Fluentd
status	95
section	TBD
keywords	TBD

Fluentd is a data collector used by many organizations such as Amazon, Microsoft, and Google[1048]. It is open source and available on GitHub. Fluentd creates a layer of abstraction between the source of the data and backend, known as the Unified Logging Layer. This centralized system of data collection ensures security and reliability. Logs contain important information, but due to modern data sizes, they are no longer for just human use. The purpose of the logging layer is to allow more machine reading of logs as opposed to human reading.

In the logging layer, data is converted to json, then sent to the backend. Fluentd has a plugin system for many different programs, such as Python and Node.js. There are also custom versions of Fluentd in production. Google, for example, uses their own version of fluentd as their logging layer in conjunction with Google BigQuery. The Fluentd project also includes Fluent Bit which is a data forwarding system.

4.95 FoundationBenchmarks



title	FoundationBenchmarks
status	95
section	TBD
keywords	TBD

The AWS Foundation Benchmarks [1049] project is a repository of Python scripts that can be used to evaluate your AWS account and its configuration. The project is intended to be integrated into your CloudFormation stack so that it can run the benchmarks on every iteration of your code pipeline. The benchmarks are sourced from the Center of Internet Security and it can help you find issues in your IAM, your VPC configuration, your S3 bucket permissions and many other places that are commonly left open by default or accident.

4.96 Future Grid



title	Future Grid
status	95
section	TBD
keywords	TBD

Future Grid works through four ways, which contains Connect, Configure, Deploy and Learn [1050].

From the official website we can know that, Performance is the most important things.

"Future Grid Platform processes and analyses billions of data points per day" [1050].

It can reduce the time of processing data about 90%. The platform can be connected to a wide variety of data sources. Furthermore, Future Grid Platform's provide end-to-end control.

Steve Avery said,

"Future Grid's innovative data platform delivers high-speed data processing that's ideal for IoT and Data Science environments in the Energy and Utility sector. We enjoy a strong working relationship with the Future Grid team and appreciate their deep understanding of Utilities as well as their ability to prepare relevant, affordable, ROI-focused Use-Cases for our clients" [1050].

4.97 Google Cloud Platform - Big data solutions



title	Google Cloud Platform - Big data solutions
status	95
section	TBD
keywords	TBD

The Google Big data solutions is a part of Google Cloud Platform services; it offers special service on data analyzes and other data engineering works. The users (usually companies, sometime personal user) could combine with the cloud services (see more in Google Cloud Platform - Cloud Dataproc in another section) to manage their data, database, and proceed data analyzing, machine learning and other related data engineering works in order to predict business decision. The Big data solutions offers all of the services about data management, efficient data query, and machines intelligence services, which is invaluable for commercial purpose [1051].

The functionality of the Google Big data solution is comparably advanced compare with other big data and cloud computing service; it has completed managed platform with not sophisticated structure, which support SQL and No-SQL data services. The Fast Queries of the Google Big data solution is also an advantage, since it saves users time and lower the cost of data searching [1051].

4.98 Google Cloud Platform - Cloud Dataproc

title	Google Cloud Platform - Cloud Dataproc
status	95
section	TBD
keywords	TBD

The Cloud Dataproc is a part of Google Cloud Platform services. It is a efficient, not sophisticated, and well managed cloud service for companies who running Apache Spark and Apache Hadoop clusters as cloud base. The companies can adopt their cloud system in to the Cloud Dataproc, and use Google Cloud Platform services to manage data, mine useful information from data and so on (see an example in Google Cloud Platform - Big data solutions). It combines with other services in Google Cloud Platform services to generates a complete cloud service for dealing with large amount of data size [1052].

More than the companies can adopt their own system into the Cloud Dataproc, which could be more suitable for the company to use, and also, the Cloud Dataproc also has many other development tools are available to use, such as the Google Cloud SDK, some web UI, and RESTFUL APIs. The richness of the development tools could lower a company cost significantly [1052].

4.99 Google Genomics



title	Google Genomics
status	100
section	TBD
keywords	TBD

Google Genomics is an extension to Google cloud platform. It helps life science community organize genomics information and make it available for research. Researchers are able to apply Google powerful technologies such as Google Search and Maps on genomics data. Google Genomics allows users to securely store, process and share complex genomics datasets. Multiple genome repositories data can be processed using Google Genomics within seconds. It is backed by high performance Google bigtable and Spanner technologies [1053].

Google Genomics is based on open standard from Global Alliance of Genomics and health. These open standards help in achieving high level of interoperability. It is fully integrated with Google cloud virtual machine, storage and databases. It helps analyze Genomic data in real-time with BigQuery. Users can analyze worldwide genomics data using their preferred language such as R, Python, Java etc through API. Data can be accessed and processed in transactional or batch mode. Users can use Genome Analysis Toolkit for analyzing Genomics data in batch mode. Apache spark or grid cluster may be used for faster processing of large and complex genome datasets. Organizations can use this platform for increasing their revenue. They can monetize their proprietary genomics data by hosting on Google Cloud and billing their clients on usage basis[1053].

4.100 Gephi



title	Gephi
status	95
section	TBD
keywords	TBD

Gephi [1054] is an open source software for visualization and exploration for all kind of graphs and network. It is a useful tool for data analyst and scientist to understand network and relationship. This tool is developed in Java and needs Java 1.7 or higher. It provides the capability to generate various graphs, interact with those graphs, manipulate the graph to discover the pattern. These graphs mostly consist of nodes and edges. Edges are nothing but the relationship between various nodes. Gephi has various layouts which provide graph in a different layout for the analysis purpose. Real-time visualization capability provides analysis by changing graph in real time through data filtering. Data filtering capability help reduce nodes and edges in the graph to do drill down analysis or keeping graph in human readable format. It has statistics and matrix framework which provides social network analysis and help community detection which is called as modularity. Gephi has Data Laboratory which allows us data manipulation as well as data transformation for analysis. It provides data import capability through various graph file format as well as CSV format. The export capability provided by Gephi exports graph in pdf and image format for analysis and presentation. It supports big data to some extent by processing capability of around 100k data points. It can be extended using built-in plugin center. It is supported on Windows, Mac OS X, Linux platforms [1055].

4.101 GitHub Developer



title	GitHub Developer
status	95
section	TBD
keywords	TBD

GitHub is a software management platform that offers free and fee-based services that can be used for managing source code for software projects. However, GitHub also offers an API through its Developer site. The API can be used to analyze a large body of data that is stored at GitHub. The data can be used to provide insight into trending software technologies, data sources that pertain to non-software management domains. For example, the GitHub resource OpenRefine [1056] is a reference to a variety of data sources that are open for public use [1057].

4.102 Apache Gobblin



title	Apache Gobblin
status	95
section	TBD
keywords	TBD

As the amount of data increases and its sources become numerous, it gets difficult to integrate this data to solve a specific problem. Apache Gobblin [1058] is a distributed data integration framework that allows users to build different data integration applications, usually as separate jobs which are executed with the help of a scheduler [1059]. Gobblin can be deployed in a stand alone manner and also supports deployment on a Hadoop, Apache Mesos or Amazon Elastic Cloud cluster [1058].

Currently Gobblin deployments run independently of each other and there is no central management or orchestration. However, efforts are being made to develop Gobblin-as-a-Service which would manage data integration jobs on any mode of Gobblin deployment [1059].

4.103 Google App Engine



title	Google App Engine
status	95
section	TBD
keywords	TBD

Google App Engine, generally called App Engine is a Platform as a service cloud solution (PaaS). It lets you build and run applications on Google's cloud infrastructure. In this platform the developer does not have to worry about infrastructure such as database administration, server configurations and load balancing which is done by google. Developers only job is to develop source codes. It claims to be highly scalable as it can automatically increases capacity depending upon the workloads [1060].

Applications in App Engine can be run in either Flexible or Standard Environment or both can be used at the same time [1061]. Automatic scaling of apps, user customization of runtime (Eclipse Jetty 9, Python 2.7 and Python 3.6, Node.js, Ruby, PHP, .NET core, and Go), operating system and even CPU memory are some of the features of App Engine Flexible environment [1061]. Applications in Flexible environment run in Docker containers on Google Compute virtual machines. While in App Engine Standard Environment application instances are run in sandbox with prespecified runtime environment of supported language (Python 2.7, Java 7, Java 8, PHP 5.5 and GO 1.8, 1.6) [1062]. This means if source code uses Python then its instances are run in Python runtime.

4.104 Google BigQuery



title	Google BigQuery
status	95
section	TBD
keywords	TBD

Google BigQuery is a cloud-based big data analytics web service for processing very large read-only data sets. BigQuery can analyze data on the order of billions of rows, using a SQL-like syntax. It runs on the Google Cloud Storage infrastructure and can be accessed with a REST-oriented application program interface (API) [1063].

BigQuery enables the creation of a logical data warehouse over managed, columnar storage as well as data from object storage, and spreadsheets. It also allows the capture and analysis of data in real-time using its powerful streaming ingestion [1064].

4.105 Cloud Bigtable



title	Cloud Bigtable
status	95
section	TBD
keywords	TBD

Cloud Bigtable is googles NoSQL big data storage service. This is currently used by google itself for their own services like search engine, Gmail, Maps. It is highly scalable big database with low latency and high throughput achieved by distributed computing. Using this database makes it easy to integrate the database with other cloud services provided from google such as cloud dataflow. Also it allows to integrate with big data tools like Hadoop. It can achieve high performance of millions of transactions per second [1065].

4.106 Google Compute Engine



title	Google Compute Engine
status	95
section	TBD
keywords	TBD

As an infrastructure as a service Google Compute Engine (GCE) provides scalable, high performance virtual machines to their clients [1066]. Its virtual machines vary in CPU and RAM configurations and Linux distributions depending on clients' need. Network storage are attached to virtual machines are attached as persistent disks. Each of these disks' size can be upto 64TB and they are automatically resized based on demands [1066]. This feature of GCE's virtual machines makes it scalable and reliable. Another feature of GCE includes its global load balancing technology which allows distribution of multiple instances across different region [1066]. It provides a platform to connect with other virtual machines to form a cluster or connect to other data centers or other google services [1066]. It can be managed through RestFul API or command line interface or web console. It claims to be cost effective and environmentally friendly compared to its competitors like Amazon Web Services.

4.107 Google Docs



title	Google Docs
status	95
section	TBD
keywords	TBD

Google Docs is a free office suite offered by Google [1067]. It is a part of Google Drive, therefore cloud syncing is native. Google Docs is a collaborative tool for creating and editing documents in real time. It allows for multiple users to edit the same file. While it has apps in Android and iOS it only has PC programs on Chrome Books.

4.108 Google Firebase



title	Google Firebase
status	95
section	TBD
keywords	TBD

Google Firebase is an Commercial cloud based platform that provides users with the capability to build mobile applications for various platforms such as Android/ios with several language support. The cloud service also provides solutions for hosting the app/database, monitoring and all necessary functionalities. The complimentary products available for the mobile applications built through the platform enable for usage Analytics, predictions, ad campaign and a ton of other development features. This platform is also available free of cost with limited functionalities for individual users.

4.109 Google Load Balancing



title	Google Load Balancing
status	95
section	TBD
keywords	TBD

Google Cloud Load Balancing is a high performance, scalable load balancing, which has state of the art auto scaling feature which distributes traffic intelligently such that application still have resources in spite of huge increase in traffic and it does not require pre-warming, as it quickly reaches from zero to full-throttle. Google Load balancing support different flavors such as HTTP, TCP/SSL and UDP Load Balancing [1068]. Also the new UI enables users to integrate any of these flavors easily through a single interface.

4.110 Google Stackdriver



title	Google Stackdriver
status	95
section	TBD
keywords	TBD

With increasing cloud-based applications it is hard for Devops engineers to keep track of performance, availability and issues associated with these applications. Google Stackdriver is a powerful service for monitoring, logging, and diagnostics. It support applications deployed on Google Cloud Platform,Amazon Web Services, and both combined [1069].

Stackdriver provides a wide variety of features such metrics, dashboards, alerting, log management, reporting, and tracing capabilities, which ultimately enables users to find and fix issues faster[1069].

4.111 Apache Gossip



title	Apache Gossip
status	95
section	TBD
keywords	TBD

Many of the applications that are cloud based and require a huge amount of computation or data storage in the back end, use cloud based clusters. When different nodes in a cluster rely on services provided by other nodes, it becomes important that each node has information about the others, to avoid failures. Apache Gossip is a protocol that provides a method that allows nodes to form a peer-to-peer network and allows them to discover other nodes and check the liveliness of the cluster [1070].

The name arises from the family of protocols known as gossip protocols or epidemic protocols that disseminate information in a manner similar to how gossip spreads in a community. Each node periodically selects any other node at random and shares the information it has, thus eliminating the need to broadcast from every node to every other node in the cluster [1071].

4.112 GraphQL



title	GraphQL
status	95
section	TBD
keywords	TBD

Data services have been an important component in the evolution of the information age. In the early 2000s data-based web services relied heavily on structured data formats like Extensible Markup Language (XML). Other data formats also included raw or plain text, or serialized data objects. Whether for public or private use, in most cases there was and has been a necessity for documenting data services. In other words, web service consumers need to know what data is available and how to query that data. In the case of XML, web-services were typically developed with Web Services Description Language (WSDL) [1072] as part of the service.

Over time, challenging issues related to using XML as a data delivery format emerged. The JSON [1073] data format in conjunction with the REST [1074] architecture emerged as an alternative to XML and SOAP. However, REST also has challenges in terms of documenting the services and data available in this type of REST architecture.

“GraphQL has emerged as query language that can reside on top of the REST architecture and address many of the issues associated with using XML/SOAP and JSON/REST” [1075].

4.113 AWS Greengrass



title	AWS Greengrass
status	95
section	TBD
keywords	TBD

AWS Greengrass [1076] is a product that allows a developer to create serverless code residing on AWS that can then be run locally on your devices. Those devices are then able to act locally on the data it creates while being able to utilize the cloud for handling the infrastructure. Devices can still communicate with each other even when unable to connect to the Internet. Filters can also be added locally to each device allowing the user to be able to control what data is being sent back to the cloud. It can help users create IoT solutions which allow connectivity between many different types of devices and the cloud simultaneously. Greengrass supports many different programming languages. One of its core features is machine learning inference, which allows a device to directly perform inference, or applying an already trained and optimized model to new data, before sending the appropriately filtered data back to the cloud.

4.114 H2O



title	H2O
status	100
section	TBD
keywords	TBD

H2O is an open source platform. It can execute highly advanced and complex machine learning algorithms in faster and scalable way. Complex models running on H2O platform give results faster regardless of size, format and place of data. H2O serializes large amount of data stored in nodes and clusters rapidly. Data processing is done in memory thus providing faster response [1077].

It uses fine grain parallelism technique for processing of distributed data. This helps H2O achieve 100 times faster speed as compared to traditional mapreduce. H2O4GPU, Sparkling Water and Driverless AI are some popular H2O products. H2O4GPU enables usage of high performance GPUs while running complex machine learning algorithm. Sparkling Water enables usage of Spark in transformation and mapping of data. And, Driverless AI is an automated machine learning platform to create artificial intelligence products and services[1077].

H2O platform is being used for machine learning and AI related research across industries. Cisco, PayPal, Comcast, Macy and Booking.com are some of the companies that uses H2O platform[1077]. H2O offers web based GUI workflow for creating machine learning models. Machine learning pipeline steps can be easily created in GUI workflow using drag and drop components. H2O platform supports mutiple languages. Developers can write machine learning code in their preferred progamming languages such as python , scala, Java, R etc.

4.115 Apache Hadoop



title	Apache Hadoop
status	95
section	TBD
keywords	TBD

The Apache Hadoop [1078] is an open-source software designed for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework using simple programming models that allows for the distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The library is designed to detect and handle failures at the application layer rather than rely on hardware to deliver high-availability. Therefore, each of the computers in the cluster may be prone to failures because it delivers a highly-available service on top of a cluster of computers. The project includes these modules: 1. Hadoop Common: The common utilities that support the other Hadoop modules. 2. Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data. 3. Hadoop YARN: A framework for job scheduling and cluster resource management. 4. Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

4.116 HBase



title	HBase
status	95
section	TBD
keywords	TBD

Apache HBase [1079] is a distributed, scalable, big data store, Hadoop database. You can use Apache HBase when you need random, realtime read/write access to your Big Data. The goal of HBase is hosting of very large tables – billions of rows X millions of columns – atop clusters of commodity hardware. Apache HBase is modeled after Google's Bigtable, which is a Distributed Storage System for Structured Data. HBase is an open source, non-relational distributed database that mimics Google's Bigtable and is written in Java. It was developed as part of the Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing Hadoop with similar Bigtable functionality. HBase provides compression, in-memory operations, and Bloom filters for each column listed in the original Bigtable file. HBase does not directly replace the traditional SQL database, but the Apache Phoenix project provides SQL layers for HBase and JDBC drivers that can be integrated with a variety of analytics and business intelligence applications.

4.117 HCatalog



title	HCatalog
status	95
section	TBD
keywords	TBD

HCatalog, which was originally known as Howl, is a component shipped with Hive that manages storage and tables. Its purpose is to simplify data storage and retrieval by providing a shared schema and data type mechanism between Hive, Pig, and MapReduce and the formats in which a Hadoop serializer-deserializer can be written (ORC, RCFile, CSV, JSON, and SequenceFile.) Custom formats can be added as well. A REST API called WebHCat (originally Templeton) is also available [1080].

4.118 HPCC Systems



title	HPCC Systems
status	95
section	TBD
keywords	TBD

HPCC (High Performance computer cluster) systems are open source tool which offers the BigData related services. HPCC contains tools that deal with complex data structure and large scale of data amount. It is a powerful open source tool for data analyze, especially for the significant size of data. The functionalities of HPCC such as fast querying to different databases, , data visualization, and data management have good reputation to many users [1081].

One of the properties of the HPCC system is easy-to-used; it is easy to learn from a developer side, and it also contains exhaustive resources for a beginner to learn. For example, the HPCC has free training and completed documentation for the new user, and even user has some unexperienced issue which is hard to solve, the development community can also help the user to learn. Despite it is easy to use, the powerful computing mechanism of the system and the massive cloud computing platform bring the HPCC system supercomputing capability [1081].

4.119 Hue



title	Hue
status	95
section	TBD
keywords	TBD

Hue or Hadoop User Interface is an open source tool licensed under Apache v2 license. It sits on top of the data at the visualization layer and provides a graphical user interface to operate and develop applications for performing self-service data analytics. The latest version of Hue available is v4.1.0, released October 4th 2017 and can be downloaded from the website [1082]. Hue is an open source Analytics Workbench for browsing, querying and visualizing data. Hue works well for a variety of technologies in the Hadoop ecosystem such as Hive, Impala, Pig, MapReduce, Spark. Query tool works with SparkSQL, Solr SQL and Phoenix. Further it works well with RDBMS such as Oracle and MySQL [1083]. Some of the applications of Hue are analytics dashboards, job scheduling, workflows, it also serves as an interface for jobs, HDFS, S3 files, SQL Tables, Indexes, Git files, Sentry permissions, Sqoop and more [1083]. Hue is available in all major Hadoop distributions such as Cloudera, Hortonworks, MapR and AWS [1083].

4.120 Hyperledger Burrow



title	Hyperledger Burrow
status	95
section	TBD
keywords	TBD

Hyperledger Burrow [1084] is an open sourced smart-contract interpreter which was built to meet the requirements of the Ethereum Virtual Machine. The Ethereum network has begun to see growth in the enterprise sector; with well-known companies such as JP Morgan, Microsoft, Accenture and BP all recently joining the Enterprise Ethereum Alliance. The importance around interpreting smart contracts created by Ethereum cannot be understated, as Ethereum has gained a lot of traction and credibility within the Cryptocurrency community and currently at the time of writing has a market cap of 82 billion USD only second to Bitcoin. Because of this widespread adoption of Ethereum, one of Burrow's claims to fame is that it is the only Apache-licensed Ethereum VM implementations on the market. [1085]

citation wrongly placed

4.121 Hyperledger Fabric



title	Hyperledger Fabric
status	95
section	TBD
keywords	TBD

Hyperledger Fabric is one of the oldest and most well known of all the Linux foundation Hyperledger projects. Initially created by IBM and Digital Asset , it's intent was to be a foundation for developing distributed ledger applications. Some of the key features sited by the team are, Channels for sharing confidential information, Ordering Service delivers transactions consistently to peers in the network, Endorsement policies for transactions ,CouchDB world state supports wide range of queries, Bring-your-own Membership Service Provider(MSP). [1086]

With many companies contributing to the growth of the platform, over 159 engineers from 28 different organizations, there is a promising future for the platform as a variety of businesses begin to explore building products with Fabric. As stated by Behlendorf the number of projects already being built is in high hundreds to low thousands. [1087]. As distributed ledger technology continues to grow, the willingness for enterprises across differing industry/sectors to contribute to this open source platform is key to it's success.

4.122 Hyperledger Indy



title	Hyperledger Indy
status	95
section	TBD
keywords	TBD

Another one of the newer developments from Hyperledger, Hyperledger Indy is all about providing independent digital identities across blockchains and distributed ledgers. It is a decentralized identity system and its advantage is that identity management is its sole focus. As Phillip J. Windley, Ph.D., Chair, Sovrin Foundation states,

“Many have proposed distributed ledger technology as a solution, however building decentralized identity on top of distributed ledgers that were designed to support something else (cryptocurrency or smart contracts, for example) leads to compromises and short-cuts” [1088].

This will allow people to securely, quickly and easily share their authenticated identity with the groups and organizations of their choosing while providing those organizations with the peace of mind of knowing who they are dealing with.

As Behlendorf states,

“Instead of being an entry in a giant data base, you have your data and deal programmatically with different organizations who want to check your identity. And companies don’t have to store so much personal data. They can store a pointer to the identity” [1084].

4.123 Hyperledger Iroha



title	Hyperledger Iroha
status	95
section	TBD
keywords	TBD

Hyperledger Iroha is an open source, mobile focused blockchain platform. The Japanese startup, Soramitsu in partnership with Hitachi started the initiative to create a mobile friendly blockchain architecture. As one of the new, up and coming Hyperledger projects it focuses on being simple and easy to include in projects and was implemented in C++ which allows it to,

“perform well with any small data projects and focused use cases.” [1084].

As stated by the Linux Foundation,

“Hyperledger Iroha is designed to be simple and easy to incorporate into infrastructural projects that require distributed ledger technology. It features a simple construction, modern, domain-driven C++ design, emphasis on mobile application development and a new, chain-based Byzantine Fault Tolerant consensus algorithm, called Sumeragi” [1089].

4.124 Hyperledger Sawtooth



title	Hyperledger Sawtooth
status	95
section	TBD
keywords	TBD

Hyperledger Sawtooth is an open source, blockchain platform which can be used to build distributed ledger applications. Its main application is to simplify the development of blockchain applications by isolating the core system from the application domain.

"This allows for developers to quickly and easily develop and deploy applications with custom tailored business rules in some of the more common languages" [1090].

Some of the core features that make Hyperledger Sawtooth a unique and interesting distributed ledger technology:

"On-chain governance - Utilize smart contracts to vote on blockchain configuration settings such as the allowed participants and smart contracts. Advanced transaction execution engine - Process transactions in parallel to accelerate block creation and validation. Support for Ethereum - Run solidity smart contracts and integrate with Ethereum tooling. Dynamic consensus - Upgrade or swap the blockchain consensus protocol on the fly as your network grows, enabling the integration of more scalable algorithms as they are available. Broad language support - Program smart contracts in your preferred language, with support including Go, JavaScript, Python and more." [1091]

4.125 IBM Big Replicate



title	IBM Big Replicate
status	95
section	TBD
keywords	TBD

To make Hadoop deployment enterprise-class, easy data replication is required to support critical business applications that depend on Hadoop. Keeping this in mind, IBM created IBM Big Replicate which does class replication for Hadoop and object store.

The main features of the product include continuous availability, high performance with guaranteed data consistency. Also, it replicates large amounts of data from lab to production environment, from production to disaster recovery sites. These replications are governed by the business rules set up. This technology replicates data as the data streams in. Thus, it reduces dependency on completion of file operation i.e., closing of file before data can be transferred. It offers replication in a flexible way by handling various Hadoop distributions and versions. Additionally, for each cluster, multiple IBM Big Replicate can be deployed as proxy servers to add resilience. Users can access Hadoop Distributed File System using Big Replicate via the standard HDFS URI [1092].

4.126 IBM Cloud



title	IBM Cloud
status	95
section	TBD
keywords	TBD

In 2017, IBM fully committed to cloud computing. IBM BlueMix is now IBM Cloud. The changes go far beyond the name. The new platform gives IBM a new, singular way to engage customers [1093]. Now, services are available on public or private clouds, with added capabilities, including database, artificial intelligence, and blockchain[1093]. In IBM Cloud, many popular services and applications are available, with public or private access. No cloud presence is possible without a strong network. IBM has included an industry leading level of network capabilities. To ensure security, access, and redundancy, IBM operates 60 data centers [1093]. The IBM Cloud connects data science and other tools, such as VMware, SAP, Spark, Jupyter, R, and many others [1093]. Both open source and proprietary applications and services are part of IBM Cloud. As an industry leader in blockchain, IBM's use of the technology is featured in IBM Cloud. Blockchain is becoming the most known product IBM offers, and it is a major component of IBM Cloud[1093].

4.127 IBM Db2 Big Sql



title	IBM Db2 Big Sql
status	95
section	TBD
keywords	TBD

citation labels do not have spaces

IBM Db2 Big Sql facilitates operations like accessing data, querying data and analysing data across data warehouses and also Hadoop.

It is a well formed hybrid engine that lets you get data by querying Hadoop using SQL. It gives you the flexibility of having a single database connection or make queries to different data sources such as

“HDFS, RDBMS, NoSql databases, object stores and WebHDFS” [1094].

One of the most important feature of this Big Sql is that it provides low latency.

This makes data retrieval easier in complex business systems. It also provides high performance, security, SQL compatibility and federation capabilities to your data warehouses.

It enables short, rapid queries that facilitates searching by key word or key ranges. It uses HBase for operations such as point queries and rapid insert. Workloads can be updated and deleted via this Hbase. To make use of easier and faster data processing in Apache Spark, it can be integrated with Spark [1094].

4.128 ID2020



title	ID2020
status	95
section	TBD
keywords	TBD

ID2020 is a new software tool sponsored by the United Nations. It will let millions of refugees and other without documents whip out a phone or other device to quickly show who they are and where they came from. The tool, developed in part by Microsoft and Accenture, combines biometric data (like a fingerprint or an iris scan) and a new form of record-keeping technology, known as the blockchain, to create a permanent identity.

The rapid proliferation of smart devices globally, combined with ever-increasing computing power and rapidly expanding broadband coverage, enables new methods of registration and facilitates ongoing interaction between individuals and their identity data.

"An estimated 1.1 billion people, including many millions of children, women and refugees, globally lack any form of officially recognized identification. Without an identity, individuals are often invisible - unable to vote, access healthcare, open a bank account, or receive an education - and bear higher risk for trafficking. Without accurate population data, public and private organizations struggle to broadly and accurately deliver the most basic human services" [1095].

Most people associate cryptocurrency with blockchain. But blockchain obviously has more to offer. This is an example where the blockchain technology has a significant social impact outside of the

usual cryptocurrency. It is so revolutionizing it could replace the data storage that we traditionally use for big data. Blockchain still has a lot of untapped potential.

4.129 IICS



title	IICS
status	95
section	TBD
keywords	TBD

Informatica provides various products in data integration and data warehousing domain. Informatica provide on premise products for Big Data, Data Integration, Data Quality, Data Security and Master Data Management. Informatica also provide cloud products for Integration Cloud, Data Quality, Governance Cloud, Master Data Management Cloud, Data Security Cloud and Data As A Service. All the cloud products comes under Informatica Integration Cloud Services (IICS). with the cloud approach customer need not to worry about the patching, high availability of the servers, upgrade etc. IICS is built on microservices architecture and modern user interfaces and provide complete end-to-end data management approach. IICS provide New and Modern User Interface Experience, Template Driven Development, Enterprise Orchestration, File Mass Ingestion, Integrated Asset Management and APIs that enable Continuous Delivery. Customer can focus on the logic of the data processing and all the infrastructure related activities will be taken care by cloud [1096].

4.130 Instabug



title	Instabug
status	95
section	TBD
keywords	TBD

Instabug is a cloud service provider which provides in-app feedback, user surveys, bug reporting, and crash reporting for mobile applications. The platform was founded in 2012 and as of 2017, Instabug has been plugged in over 800 million devices including most of the top 100 apps in Android, iOS, Cordova, Ionic, Xamarin, and web application markets[1097]. As the Instabug [1097] claims most of the top apps in the world rely on Instabug for beta testing, user engagement and crash reporting because of it is reliable, and easy to integrate (it comes with customizable SDK). Instabug is well known for the customizable Shake to Send feature on the mobile app to invoke the bug reporting, annotated screenshots, voice note or a screen recording to better describe the bug to provide a descriptive report on the developer side without interrupting the user experience [1098]. Yahoo, soundCloud, paypal, Lyft, Buzzfeed, Kik and Nextdoor are few famous applications which uses Instabug for bug and crash reporting[1097].

4.131 Intel Cloud Finder



title	Intel Cloud Finder
status	95
section	TBD
keywords	TBD

Intel Cloud finder is an enterprise level solution for choosing cloud service provider. It helps the customers seeking help for cloud service providers. It also provides a very good resource for people looking for a very good performance with decent amount of security in the cloud. They state that

"In the domain of Intel Cloud Technology we have Intel Advanced vector expansions, Intel Turbo boost technology and Intel Xeon processor but on the security side we have Intel trusted execution technology hardware-based protection for the cloud, ensuring a secure foundation and protecting applications against malware, malicious software, and other attacks" [1099]

looking forward to using this tool for selecting cloud services that would satisfy personal cloud requirements.

Obviosly we need to wait until Intel makes this tool open to public as this looks to be proprietary tool.

4.132 Jaspersoft



title	Jaspersoft
status	95
section	TBD
keywords	TBD

Jaspersoft is a Business Intelligence (BI) platform that provides its customers with highly interactive reports, analytics and dashboards. This tool has been downloaded more than 14.5 million times which makes it the most cost-effective, flexible and widely used platforms in the world [1100]. It is embedded with data visualization, analytics, and reporting capabilities that allows its customers to gain insight from various data sources and enables them to make better decisions [1100]. This reporting software is designed to take input from one or more data sources, including Big Data [1101], NoSQL [1102], JDBC [1103], XML [1104], JSON [1105], CSV [1106], Hibernate [1107], POJO [1108] and Web Services and present it in an easy-to-read, highly interactive format for business users [1109]. The print ready, interactive reports and dashboards enables organizations to interact with their data both inside or outside their organizations which also enables for faster business decision-making [1109]. This tool provides a centralized repository, in which a customer can store user-profiles, dashboards, analytic views, reports, and more. It supports thousands of users and is designed for small, medium and big organizations including enterprises [1100]. It has a Java-based reporting library JasperReports [1100], that provides pixel-perfect documents and generates ad hoc based reports for the web, the printer or mobile device [1109]. JasperReports is an open source reporting tool that can be used in any Java-enabled applications [1110]. Jaspersoft is a gold partner with MySQL and was recently acquired by TIBCO in April 28, 2014 [1110].

4.133 JavaScript



title	JavaScript
status	95
section	TBD
keywords	TBD

JavaScript is a ubiquitous programming language with an enormous number of uses. It is best known as the language used to create web pages[1111]. However, its uses are far more than would appear. JavaScript is a flexible language that can function as object oriented, as well as procedural[1111]. JavaScript has had a storied past[1111]. Because of the universal adoption of JavaScript, it has a bright future. JavaScript is not confined to rendering web content. As an object-oriented language, JavaScript has unlimited potential as a programming platform[1111].

4.134 Jelastic



title	Jelastic
status	95
section	TBD
keywords	TBD

Jelastic is a cloud service provider which combines platform as a service and container as a service in a single package.

The main features include built-in metering, monitoring and troubleshooting tools. It is available as a public, private, hybrid and multi-cloud application. It can manage multi tenant Docker containers to native ecosystem. It facilitates live migration of workloads across various regions and various clouds with zero downtime. This makes the system highly reliable during migration. All the resources from different cloud environment can be accessed using a single panel. It also supports microservices and legacy application with absolutely no code changes. It provides integration with Git, SVN and CI/CD tools and services. It enables scripting to automate processes and events in the cloud.

In terms of languages, it supports various languages such as Java, PHP, Ruby, Node.js, Python, .NET and Go. Additionally, it supports virtualization technologies like Docker and Virtuozzo. It also supports a wide range of databases such as MySQL, MariaDB, Percona, PostgreSQL, Redis, Neo4j, MongoDB, Cassandra, CouchDB and OrientDB [1112].

4.135 JMP



title	JMP
status	95
section	TBD
keywords	TBD

JMP, commonly referred to as Jump, is an enterprise level statistical analysis tool developed by SAS. The JMP software package is designed to handle every data-involved stage from the initial acquisition of data to the final presentation of findings. JMP was first released in 1989 and has been designed ever since to provide a visual-centric interface where the user can analyze, manipulate, and format data. JMP is capable of complex analysis and machine learning techniques and can provide the user the back-end software code generated to produce the visualized results in a variety of common statistical languages or applications such as Python, R, Matlab, SAS, and others [1113]. A single JMP license is available for 1,785 USD [1114]. JMP Pro is an even more capable version of JMP with more advanced analytics and predictive modeling with cross-validation-available for 14,900 USD [1115].

4.136 Kafka



title	Kafka
status	95
section	TBD
keywords	TBD

Kafka is an open source distributed streaming platform that belongs to the Apache Hadoop family. It is mainly used in data ingestion and building real time data pipelines. Since Kafka is fault tolerant, scalable and efficient it is in production today in thousands of companies. Kafka also works as messaging system based on the publisher-subscriber model where the publisher produces data and the subscriber consumes the data. It can be compared to ActiveMQ in the messaging space [1116].The two main real-time streaming applications of Kafka are, building data pipe-lines to transport data between systems or applications, and building applications that transform or react to stream of data. A few use cases of Kafka are website activity tracking, messaging and log aggregation [1116].

4.137 Keras



title	Keras
status	95
section	TBD
keywords	TBD

“Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK or Theano” [1117].

Keras is used in deep learning community to build neural nets with easy to use syntax. There are other technologies/software packages like TensorFlow, Pytorch which are also used to build neural networks, but the syntax and documentation is long and difficult for a beginner in deep learning community. Keras kind of acts as a wrapper around TensorFlow with easy to use syntax but not compromising on different tools/options within building neural network.

Using Keras, we can build normal Dense ANNs, Convolutional Neural Nets, Optimize our neural network by specifying loss function and optimization method, do regularization of neural network with techniques such as dropout, etc..

4.138 KNIME



title	KNIME
status	95
section	TBD
keywords	TBD

KNIME Analytics Platform [1118] (otherwise known as Konstanz Information Miner) is an open source platform focused on including machine learning components, data mining, analytics, and reporting; all executed through an interactive visual interface displaying modules linked together through a pipeline workflow. These modules are made up of data analytic routines which are comprised of either R, Weka, or KNIME's own native routines. Since coding is optional through KNIME's graphical interface [1119], workflows can be created to represent the individual steps in a dataflow, and allows the user to execute these steps selectively and view their output throughout different stages of their workflow [1118].

KNIME can allow for the use of additional extinctions and plugins since it's java based and built on Eclipse. In addition KNIME supports common dbms as data sources [1118]. Extensions available to KNIME also include Big Data Extensions to utilize computational resources from sources such as Apache Hadoop, Spark, and the KNIME Server. KNIME also allows the ability to blend various data sources, such as databases, images, text files, XML, JSON, networks, and Hadoop data, to be included in the same pipeline workflow and provide data and tool blending capabilities [1119].

4.139 Kubernetes



title	Kubernetes
status	95
section	TBD
keywords	TBD

Kubernetes is an open-source platform designed to automate deploying, scaling, and operating application containers.

"The name Kubernetes originates from Greek, meaning helmsman or pilot, and is the root of governor and cybernetic" [1120].

Kubernetes is capable of scheduling and running application containers on both physical or virtual clusters. Kubernetes allows developers to design applications that are agnostic of underlying architecture and allows developers to design applications based on a container-centric infrastructure rather than host-centric infrastructure, utilizing the full advantages and benefits of containers [1121]. Kubernetes via its building blocks (primitives) provides mechanism for easily deploying, maintaining, and scaling applications. Underlying architecture of Kubernetes is meant to be loosely coupled and extensible so that it can be used across a wide variety of workloads [1122].

4.140 Kudu



title	Kudu
status	95
section	TBD
keywords	TBD

Apache Kudu was designed to fit into the Hadoop ecosystem and it serves as the storage layer that enables fast analytics on fast data [1123].

Kudu internally follows the columnar storage approach rather than storing data in rows. This columnar approach helps in efficient encoding and compression. Kudu serves as a good alternative to HDFS and Apache HBase. It works best especially with use cases that require fast analytics on fast data. It is also efficient and designed to take advantage of next generation hardware and in-memory processing [1123].

4.141 Kylin



title	Kylin
status	95
section	TBD
keywords	TBD

Apache Kylin is an Online analytical processing or OLAP engine developed specially for Big Data applications. Analysts around the world today use SQL interfaces to query data from traditional database systems, Kylin fits in well due to this as it provides SQL interface as well as OLAP capability for Hadoop ecosystem which no other tool provides today. Kylin is very efficient and can return billions of rows in minimum time. It integrates well other BI tools such as Tableau. Kylin is rich in features and provides, incremental refresh of cubes, web interface for monitoring and management and Idap integration as well. Since Kylin was specially built for Hadoop and big data applications, it depends on some of the components such as HBase, Hive and HDFS. HBase is used to store the data cube, map reduce does the cube refresh or partial refresh job and HDFS stores intermediate files during cube building process [1124].

4.142 LightGBM



title	LightGBM
status	95
section	TBD
keywords	TBD

ERROR: CITATION PLACEMENT WRONG WE CAN NOT FIGURE OUT IF THIS MEANS IT IS PROPERLY QUOTED

"A fast, distributed, high performance gradient boosting(GBDT, GBRT, GBM or MART) framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks. It is under the umbrella of the DMTK project of Microsoft" [1125].

LightGBM is used to implement gradient boosting algorithm in machine learning with the aim to do so fastly, at the same time not compromising on high performance. A couple of lines on gradient boosting is necessary in understanding the context and relevance of LightGBM. Gradient Boosting is a machine learning technique used to build both regression and classification models. It is primarily used in building decision trees. But building gradient boosting models on huge datasets(that sometimes contain more than 500,000 observations) is computationally onerous, not so efficient. LightGBM solves this problem and that is why it is gaining popularity in Machine Learning community and people are using this in their Kaggle machine learning projects.

4.143 Lingual



title	Lingual
status	95
section	TBD
keywords	TBD

Lingual is a free, open source project designed to build Big Data applications on Apache Hadoop [1126]. All dependencies are installed through Maven [1127], thereby allowing the developers to focus on simply creating the applications which makes this tool easy to use [1128]. Lingual leverages the platform support of Cascading [1129], a stand-alone open source Java application framework used for building data-intensive, enterprise Big Data applications and frameworks on Hadoop [1126]. Whether on-premise or in the cloud, Lingual is compatible with all major distributions of Hadoop [1126]. Its ANSI-standard SQL [1130] interface allows SQL users to utilize their existing SQL skills to access data locked on the Hadoop [1131] clusters, thereby allowing them to create Big data applications instantly without undergoing any new training [1126]. It also provides a JDBC [1103] driver, that can be integrated with many existing BI tools and application servers [1128]. Being ANSI-SQL compliant, Lingual enables companies to query and export data from Hadoop directly into traditional BI tools [1128]. Lingual also provides other features like a SQL shell which is an interactive SQL command interface to interact with Hadoop and a Catalog to map the database tables into Hadoop files and resources [1128]. The ability to migrate workloads on to Hadoop either through Cascading applications or with the use of legacy SQL statements, significantly reduces the computing costs [1128]. Due to its ease of creating applications using SQL, JDBC or traditional BI tools, it overcomes the barriers of integrating Hadoop with the existing data management systems enabling fast and simple Big Data application development on Apache Hadoop [1126].

4.144 LinkedIn WhereHows



title	LinkedIn WhereHows
status	95
section	TBD
keywords	TBD

LinkedIn WhereHows is an open source project carried out by the LinkedIn Data team. The project works by creating a central repository and portal for several important elements of big data systems: the processes, people, and knowledge around the data [1132]. The repository has captured the status of 50 thousand datasets, 14 thousand comments, 35 million job executions and related lineage information [1132]. WhereHows integrates with all LinkedIn data processing environments and extracts metadata before offering this piece of information through two interfaces: one is a web application which facilitates functionalities such as navigation, search, lineage visualization, annotation, discussion, and community participation; the other is an API endpoint that empowers automation of other data processes and applications [1132]. The name WhereHows comes from the two fundamental questions related to the data: where is the data, and how is it produced and consumed [1133].

4.145 Linode



title	Linode
status	95
section	TBD
keywords	TBD

Linode is a cloud service provider. It provides compute, storage and networking services on demand. Linode provides SSD Storage which is Industry-leading native SSDs for optimal performance, 40Gbit Network which is 40Gbps throughput with multiple levels of redundancy and Intel E5 Processors which are the fastest processors in the cloud market. Linode has three regions and 9 data centers across the world. These regions and data centers help in data recovery and fault tolerance in case of failures. Customers use Linode services on demand as infrastructure for their websites, web services and applications. Linode provides Two-Factor Authentication, IPv6 Support, Rescue Mode, DNS Manager, Scaling, Cloning of the configuration, Supported Distributions for various distribution images. The Linode easy interface Linode Manager allows to deploy, boot, resize and clone in just a few clicks. Linode offers different billing plans and customers can select the plan as per their requirement [1134].

4.146 Logicalglue



title	Logicalglue
status	95
section	TBD
keywords	TBD

Logicalglue [1135] is a predictive analytics software that is mostly targeted towards insurance sector. It employs fuzzy logic to generate rules which in turn derive accurate predictions. Logical glue helps in identifying which data is predictive and can be deployed in cloud. It has its API which can be integrated into business's already existing softwares. Logicalglue employs machine learning and genetic algorithms to generate outcomes. New and dynamic data can be fed to the model generated and analysis can be run in realtime. This model accurately works on complete lifecycle of a project right from customers acquisition to closure.

4.147 Lumify



title	Lumify
status	95
section	TBD
keywords	TBD

Lumify is an open source project developed at US national security contractor Altamira, with key features including big data fusion, analysis, and visualization platform. The web-based interface provides users with the ability to discover connections and explore relationships in their data via various analytic options. These options include 2D and 3D graph visualizations, full-text faceted search, dynamic histograms, interactive geographic maps, and collaborative workspaces shared in real-time [1136]. Lumify has an Open Layers-compatible mapping system which can be utilized by tools like Google Maps to display an interactive geospatial analysis of the data set. Further, Lumify was integrated with SAP's high speed HANA in-memory database and computation engine, which enables faster data retrieval and calculation speed compared to the use of conventional database system [1137]. By August 2017, Altamira's Lumify is available through both the Microsoft Azure Marketplace and Amazon AWS Marketplace. The tool can be immediately run on the Azure and AWS cloud platforms, where customers have the option to purchase a license from Altamira [1138][1139]. These cloud technologies allow for greater flexibility and usability of Lumify.

4.148 Apache Mahout



title	Apache Mahout
status	95
section	TBD
keywords	TBD

Apache Mahout, an Apache Software Foundation project, is a distributed Scala DSL based linear algebra framework designed to aid mathematicians, statisticians and data scientists in implementing their own algorithms quickly and efficiently [1140]. Initiated based on Andrew Ng et al.'s paper

"Map-Reduce for Machine Learning on Multicore" [1141], it has evolved over time to cover other general machine-learning approaches [1142].

While Apache Spark is recommended back end, where core algorithms are implemented on top of Apache Hadoop, Mahout is also extensible to other back ends and standalone implementations [1143]. While number of algorithms supported by Apache Mahout is increasing, its core algorithms primarily contain implementations for clustering, classification, and Collaborative filtering [1143].

4.149 MapBox



title	MapBox
status	95
section	TBD
keywords	TBD

MapBox is a geospatial data and location platform for multiple application forms. The MapBox platform provides services and features that can be used for storing geospatial data, cartographic map production, and web-based map interface development tools. MapBox also provides a software-as-a-service features that allow users to build base-maps that can be used through various API tools in custom applications that might require maps or geospatial data. Although, a proprietary platform MapBox offers free services and access for limited use [1144].

4.150 MariaDB



title	MariaDB
status	95
section	TBD
keywords	TBD

ERROR: TOO MANY CITATIONS

MariaDB is an open source relational database. It has 12 million users worldwide and one of the fastest growing databases in the world [1145]. It powers applications at leading companies like booking.com, Virgin Mobile, HP etc. [1145]. It was created by the original developers of MySQL [1146] and Michael Monty Widenius is the lead developer of MariaDB, also the founder of MySQL AB [1147]. It was developed to be an enhanced, drop-in replacement for MySQL [1148]. It is available in Debian [1149] and Ubuntu [1150], and is now the default database on many Linux [1151] distributions. It supports a broad set of use cases by using different storage engines for different use cases. These pluggable storage engines make MariaDB a flexible, robust and scalable database solution [1152]. Up until 10.1 version of MariaDB, it used Percona's XtraDB as the default storage engine, but from version 10.2, MariaDB uses InnoDB as the default storage engine [1153]. Additional featured storage engines in 10.2 version include MyRocks for better performance and efficiency, and Spider for scaling out with distributed storage, however these are under technical preview [1152]. MariaDB intends to maintain compatibility with MySQL [1147], but it also aims at providing a rich ecosystem of storage engines, plugins and many other tools to make it versatile for a wide variety of use cases [1148]. It is used to turn data into structured information for a wide variety of applications, ranging from banking to websites. MariaDB also supports GIS and JSON features in its latest versions [1148]. It is highly secure, reliable and trusted by the world's leading brands. It is used to support enterprise

needs from OLTP to analytics [1145].

4.151 Mesosphere



title	Mesosphere
status	95
section	TBD
keywords	TBD

Mesosphere is an Datacenter Operating Platform for data-intensive applications. It is based on the Apache Mesos kernel [1154]. It is a top-level cluster, manager, container platform and operating system [1155]. Mesosphere performs resource consolidation, resource isolation, and storage capabilities in a scalable system as it runs distributed containerized software. It is agnostic to the infrastructure level, and so can be run on either physical or virtual machines [1156]. Mesosphere incorporates (1) with Amazon AWS and Microsoft Azure.

(1): Please can you elaborate?

4.152 Metron



title	Metron
status	95
section	TBD
keywords	TBD

Apache Metron is yet another open source project that serves as a big data solution for CyberSecurity applications. It provides a framework to ingest, process and also store data such as application logs, network logs and so on so that can be analyzed by Information security teams so that they can detect anomalies and respond to cyber threats. It provides the storage solution in the form of security data lake or vault where the logs can be stored long term on cost effective storage. In addition to SIEM or Security information and event management features Metron also provides packet replay utilities which can be of immense help for the security analysts. Metron also support applying machine learning algorithms on real time data that is being ingested through continuous streams [1157]. Apache Metron caters to personnels at all levels in the CyberSecurity operations from CISO to SOC analyst to security Data Scientist [1158]. It provides a single view of the risk to CISO or Chief Information Security Officer while automatically performs that analytics so security investigator does not have to spend time on finding correlation in the data. Metron also has the capability to create incidents and can integrate with the ITSM or Information technology service management systems to provide traceability [1158].

4.153 Apache Milagro



title	Apache Milagro
status	95
section	TBD
keywords	TBD

As an increasing number of connected devices communicate data with each other, data security must be taken into account. Apache Milagro is a security framework, built for cloud based software and Internet of Things (IoT) applications, that require to be scalable [1159]. Apache milagro is a pairing-based cryptography system that distributes cryptographic operations among various entities to provide a deeper level of security as compared to monolithic certificate based systems used today [1160].

Apache milagro avoids the problems faced by certificate based systems like Public Key Infrastructure (PKI) such as single point of failure, by the use of distributed trust authorities (D-TAs), which hold a part of a client's key each and are isolated from each other. The absence of a central certificate provider means that anyone can be a D-TA the key lifecycle is a part of the crypto system itself [1160].

4.154 mLab



title	mLab
status	95
section	TBD
keywords	TBD

citation is place wrong. check how to cite

mLab is an efficient service to host MongoDB databases with fully managed cloud database services. mLab has partnered with platform-as-a-service providers and it also runs on cloud providers such as Amazon, Google, and Microsoft Azure.

The main goal of mLab is to make software developers more productive. This is achieved by providing a total package of mLab which includes managed cloud database service featuring along with automated provisioning and scaling of MongoDB Databases, backup, recovery, monitoring, web-based management tools, and expert support [1161].

4.155 MonetDB



title	MonetDB
status	95
section	TBD
keywords	TBD

MonetDB is an open-source, column oriented database system. MonetDB mainly targets being a backend database for business oriented applications. These applications create very large databases having millions of rows and hundreds of columns and MonetDB supports scalability for such systems. MonetDB comprises of three software components namely the SQL front-end, tactical-optimizers and abstract-machine kernel [1162]. In contrast with MongoDB, the primary database model for MonetDB is a Relational DBMS, but it also has additional document and key-value stores. The most notable characteristic is that MonetDB is a column-store in memory that is optimized for geo-spatial support and has JSON document support [1163]. MonetDB is developed by CWI in Netherlands. It targets Big Data and Deep Learning applications and Online Analytical Processing (OLAP) and it is widely used in the Netherlands as an analytical software for Customer Relationship Management (CRM). It can also be considered as a valuable contribution to the IT industry from the Dutch [1164].

4.156 MongoDB



title	MongoDB
status	95
section	TBD
keywords	TBD

MongoDB is document database that belongs to the NoSQL family of databases. MongoDB is free and open-source, published under the GNU Affero General Public License. It is known for its scalability and flexibility. MongoDB stores data in flexible JSON-like documents. MongoDB's HA features include automatic failover and data redundancy, this is achieved using replica set, which is nothing but group of MongoDB servers that maintain the same data. MongoDB supports sharding by distributing the data across the cluster of machines [1165].

4.157 Morpheus



title	Morpheus
status	95
section	TBD
keywords	TBD

Morpheus provides cloud and hybrid cloud solutions to improve the efficiency of continuous development and integration life cycles by focusing on devops and developer perspectives. The analytics offering of Morpheus focuses on optimizing resource allocations on VM environments, such as containers and public clouds, that distributes over multiple clouds with platform independent discovery services. The competitive edge that the Morpheus has over other VM boost up package vendors is the Analytics' ability to visualize platform wide resource consumptions. The Analytics pack uses either built in cloud APIs or specific agents to gather resource consumption information across all the platforms and does the brokerage of incoming requests to minimize the incurring resource consumption costs [1166].

Morpheus governance tool provides the ability to index, categorize and store enterprise artifacts and provide life cycle management of the artifacts. The integrated Role Based Access Control (RBAC) makes sure that the artifacts are accessible only to the authorized people and provides them with the ability to view certain aspects of the artifacts. The artifacts could be anything that comes under the hood of SOA governance [1167] and they are managed by the policies defined and uploaded by the authorized roles in the system.

4.158 Microsoft Visual Studio



title	Microsoft Visual Studio
status	95
section	TBD
keywords	TBD

Microsoft Visual Studio (MVS), community edition, is an open source

“integrated development interface IDE applicable for the development of computer programs, websites, web services, web and mobile apps” [1168].

While the interface consists of some built-in tools such as code editor, code profiler and integrated debugger, it also supports plugins depending on requirements of visual designer. Javascript, C++, XML, CSS, .NET are among some of the built -in languages for visual studio. But it supports quite a big number, that is 36 different, types of languages. Python, Ruby, Node.js and M are some of the languages available by plugins. Multiple instances with their own set of packages and specific App-id, can be run at the same time. MVS connects to windows AZURE making it portable for developers [1168].

4.159 neo4j



title	neo4j
status	95
section	TBD
keywords	TBD

Neo4j is a graph database developed by Neo4j, Inc. It is an ACID-compliant transactional database with graph storage and processing which in turn help data scientists to gain new perspectives on data. Neo4j's Graph Platform is specifically optimized to map, analyze, store and traverse networks of connected data to reveal invisible contexts and hidden relationships to help enterprises tackle challenges such as Artificial Intelligence and Machine Learning, Fraud Detection, Master Data Management [1169]. Neo4j is one of the popular Graph Databases and Cypher Query Language (CQL). Neo4j is written in Java Language. Neo4j provides a flexible simple and yet powerful data model, which can be easily changed according to the applications and industries. Neo4j provides results based on real-time data and it is highly available for large enterprise real-time applications and also it does not require complex joins to retrieve the data. Neo4j can connect to REST API to work with programming languages such as Java, Spring, Scala etc.

4.160 Neptune



title	Neptune
status	95
section	TBD
keywords	TBD

Neptune is a graph database service that was announced at the AWS Re:INVENT conference in November of 2017 [1170]. Graph databases are NoSQL databases that used graph structures to organize data [1171]. They are commonly used for social networking applications, but can be used for recommendation engines, logistics, and other applications. Amazon offers Neptune as a fully managed product. It supports Apache TinkerPop Gremlin and SPARQL open source graph APIs. One can choose Gremlin or the W3C standard Resource Description Framework model [564].

4.161 Netflix



title	Netflix
status	95
section	TBD
keywords	TBD

Netflix offers a big-data platform that has both data and services that can be used to access data as well as processing tools (algorithms) that can be used to analyze data. For example, Netflix's platform has tools that can be used to detect outliers in large data sets [1172]. Netflix's data sets can be consumed and analyzed with their platform analytical tools or the tools can be used independently on other non-Netflix data sets [1173].

4.162 Apache NiFi



title	Apache NiFi
status	95
section	TBD
keywords	TBD

Apache NiFi, which is short for NiagaraFiles, is an open source software project from the Apache Software Foundation designed to automate the flow of data between software systems [1174]. Based on the NiagaraFiles software previously developed by the NSA, Apache NiFi is part of its technology transfer program in 2014 [1174]. NiFi executes within a Java Virtual Machine with the following primary components: Web Server, Flow Controller, Extensions, FlowFile Repository, Content Repository and Provenance Repository. Since NiFi's fundamental design concepts are closely related to Flow Based Programming (FBP), some of the above components can be mapped closely to FBP terms. For example, Flow Controller and FlowFile can be related to Scheduler and Information Packet in FBP terms respectively [1175] [1176]. Apache NiFi supports scalable directed graphs of data routing, transformation, and system mediation logic, aiming at leveraging the capabilities of the underlying host system on which it is operating, especially with regard to CPU and disk [1175]. Some of the high-level capabilities and objectives of Apache NiFi include: Web-based user interface, Highly configurable, Data Provenance, Designed for extension and Secure [1177].

4.163 Node.js



title	Node.js
status	95
section	TBD
keywords	TBD

Node.js is a JavaScript runtime that provides a scalable option for network applications. Node.js requires less system resources to perform than is common [1178]. Using less system resources does allow Node.js to scale. This is ideal for large web applications, such as libraries[1178]. Generally, Node.js creates server-side applications [1178]. JavaScript language is used by Node.js, and the two often compliment each other, on opposite sides of a transaction [1178].

4.164 ODBC



title	ODBC
status	95
section	TBD
keywords	TBD

Odbc is an R package which allows connectivity to commercial databases, such as Oracle, and MS SQL Server [1179]. It also permits connection to other databases with odbc (Open Database Connectivity) hooks, however other packages simplify these connections [1015]. Because odbc is actually a thin wrapper around the c++ ODBC bindings, it is faster than any other common database connector [1016].

4.165 OneDrive



title	OneDrive
status	95
section	TBD
keywords	TBD

Microsoft's OneDrive is a data storage service [1180]. It allows accessibility to files stored in any computer connected to the web. OneDrive has sharing functions and integration with Microsoft Office. Its intended audience is both personal users and companies.

4.166 Oozie



title	Oozie
status	95
section	TBD
keywords	TBD

Apache Oozie is a workflow scheduler to manage Hadoop jobs. Workflows are defined as a collection of control flow and action nodes in a directed acyclic graph. Oozie jobs can either be triggered based on the frequency or based on when the data becomes available. Oozie is integrated with the Hadoop stack supporting several types of Hadoop jobs such as map-reduce, Pig, Hive, Sqoop, Java programs and shell scripts. Oozie is an Open Source Java Web-Application available under Apache license 2.0. Oozie is a scalable, reliable and extensible system. It is responsible for triggering the workflow actions, which in turn uses the Hadoop execution engine to actually execute the task. Hence, Oozie is able to leverage the existing Hadoop machinery for load balancing, fail-over, etc [1181].

4.167 Openchain



title	Openchain
status	95
section	TBD
keywords	TBD

Openchain is a blockchain ledger technology designed to be built in seconds and deployed on an enterprise scale for the purposes of managing asset transactions. The Openchain technology uses a distributed client-server architecture rather than the slower peer-to-peer proof of work concept originally adopted by early blockchain technologies, most notably: Bitcoin. Openchain can serve as a stand alone ledger or can be scaled as a side-ledger to existing blockchain platforms [1182]. Openchain's parent company, Coinprism, has proven technology projects adopted by multiple large-scale companies like the Open Asset technology used by NASDAQ [1183]. Openchain is capable of handling any type of digital asset such as gift cards, legal documents, contracts, coins or tokens, or client-specific asset data and this is done through the enacting company serving as its own administrator by creating and enforcing it's own trust-approver validation hierarchy [1184].

4.168 OpenDaylight



title	OpenDaylight
status	95
section	TBD
keywords	TBD

OpenDaylight is an open source Software Defined Networks (SDN) controller[1185]. SDN separates network control logic from physical networking equipment. The result is that networking equipment is programmable like other computing platforms. SDN facilitates Network Functions Virtualization (NFV), allowing virtual network services (switching, virtualized appliances, and virtualized applications) to be deployed without having to deploy specialized physical devices[1186]. The OpenDaylight project was founded by a group of large tech companies, including Cisco, Citrix, Ericsson, HP, IBM, Microsoft, NEC, Red Hat, and VMware. Microsoft and VMware have since left the project[1187].

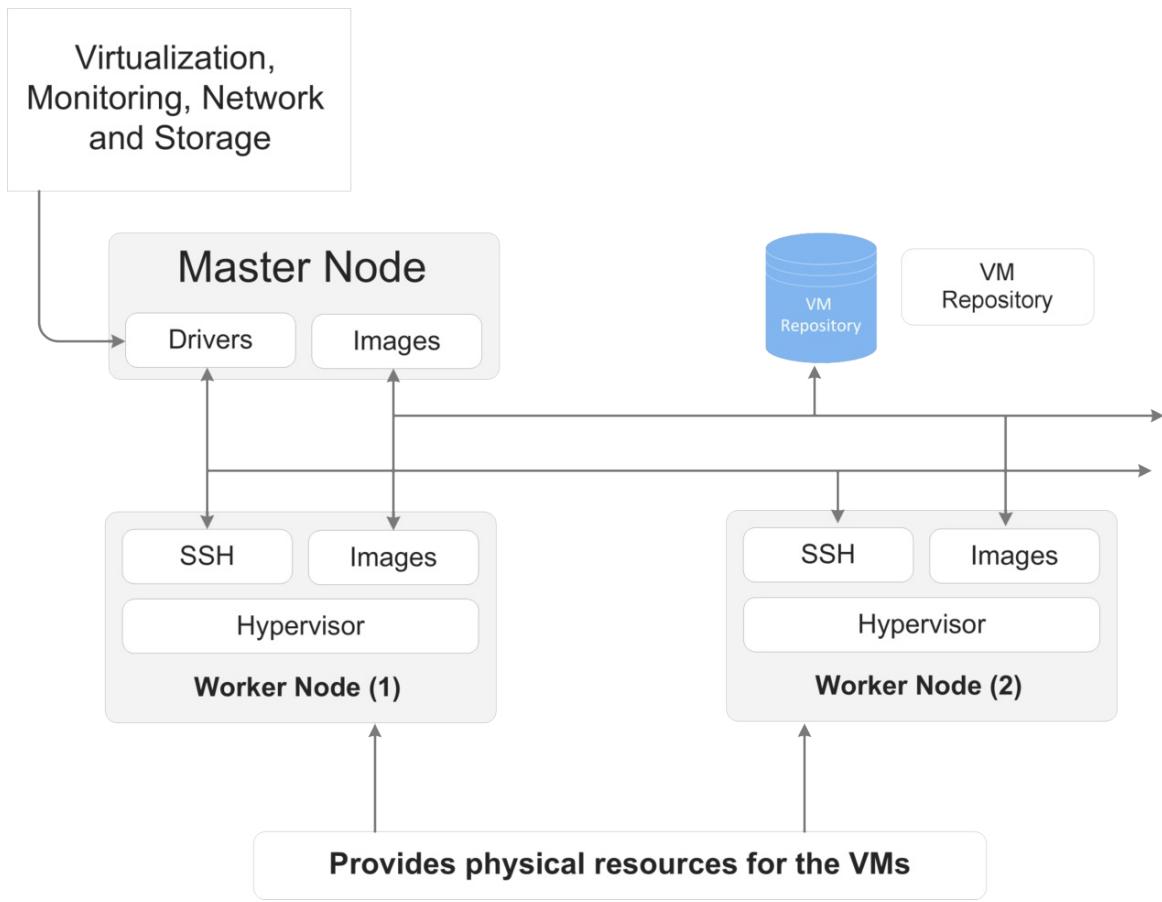
4.169 OpenNebula



title	OpenNebula
status	95
section	TBD
keywords	TBD

OpenNebula is a useful opensource that enables seamless management and control of different cloud systems. The tools can be used for a cloud implementations to virtualize data centers and also to obtain solution for cloud infrastructure. Opennebula can be adopted on top of an existing cloud setup. OpenNebula project started on 2005 and currently the product is available as an open-source under Apache license.

"The toolkit includes features for integration, management, scalability, security and accounting. It also claims standardization, interoperability and portability, providing cloud users and administrators with a choice of several cloud interfaces (Amazon EC2 Query, OGF Open Cloud Computing Interface and vCloud) and hypervisors (Xen, KVM and VMware), and can accommodate multiple hardware and software combinations in a data center" [1188].



OpenNebula Deployment Model [1189]

The OpenNebula deployment needs (1) A client node (2) A hypervisor (3) A data storage system (4) Physical network. The deployment model is depicted in [\[F:opennebula\]](#). Due to its long steady growth, the tool is being used by customers in various industries ranging from telecom to education. The wide range of customer base is helpful in providing a solid support system to the new and existing users as well as continuous feedback becomes vital in the research and growth of the project.

4.170 OpenNN



title	OpenNN
status	100
section	TBD
keywords	deep learning, OpenNN

OpenNN is a collection of functions developed for research work in field of Deep learning and deployment of neural network. OpenNN is an open source library. Development of OpenNN libraries were using C++ programming language, which provided a benefit of low memory usage and high performance with less computation time.

With multiprocessing, shared memory and Graphics processing units accelerations used for computing multiple calculations simultaneously, which provides programs to have CPU parallelism and result in better and faster computation. OpenNN provides algorithms for data mining in form of libraries and functions, which are utilized with data mining and predictive modeling [1190].

OpenNN libraries and functions are integrated with analytical tool, such as Neural Designer[1191] which provides a UI interface for users. This provide the ease of visualization for the users, to do any data entry tasks or once the model is created and output is generated, can provide results interpretation[1190].

With OpenNN, the neural network is establishing multilayer perceptron, which describes the count of neurons and their respective connectivity. OpenNN model can be enhanced with scaling or unscaling, where the layer can exist with basic statistics of mean, standard deviations, minimum or maximum values of the input or output variables[1192].

4.171 Open Refine



title	Open Refine
status	95
section	TBD
keywords	TBD

OpenRefine is a useful open source that is used for data visualization and analysis. Its predominantly used for cleaning messy data and transformation of data from one format to other for ease of clarity. OpenRefine was formerly known as GoogleRefine. The tool is also used for fetching data from websites and data organization. It can import data from CSV, TSV, Excel, XML etc. It is written in Java. It works with data in tabular format like in relational data. The tool has a user interface that is available to be downloaded.

“Once you get used to which commands do what, this is a powerful tool for data manipulation and analysis that strikes a good balance between functionality and ease of use” [1193].

4.172 openVZ



title	openVZ
status	95
section	TBD
keywords	TBD

OpenVZ is a container-based virtualization for Linux [623]. OpenVZ is mainly composed of three parts: the kernel, an set of tools and operating system templates.

OpenVZ applies to many places, including server consolidation, hosting, development and testing, security and education [1194]. If you have a lot of Linux servers that are not being fully utilized, OpenVZ can be used to integrate them into a few (or possibly one) physical machines. However, OpenVZ is completely command-oriented and currently does not include any GUI-based applications. It may not be suitable for those who are afraid of a shell prompt.

4.173 Oracle Big Data Cloud Service



title	Oracle Big Data Cloud Service
status	95
section	TBD
keywords	TBD

Oracle Big Data Cloud Service is an automated service that provides a high-powered environment tailor-made for advancing businesses' analytical capabilities. With automated lifecycle management and one-click security, Big Data Cloud Service is designed to optimally and securely run a wide variety of big data workloads and technologies while simplifying operations [1195] Oracle Big Data Cloud Machine use connectors to seamlessly integrate with other Oracle services such as Oracle R Advanced Analytics for Hadoop (this enables the development of models using R that run in parallel on Hadoop, and accelerated using Spark), Oracle Data Integrator (this enables the transformation and enrichment of data within an enterprise big data clusters), Oracle SQL Connector for Hadoop and Oracle Loader for Hadoop (this allows the integration with data in online or offline modes), Big Data Spatial and Graph (this enables processing and enriching geospatial data), and Oracle R Advanced Analytics for Hadoop (this enables building models using R that run in parallel on Hadoop, and accelerated using Spark) [1195]

4.174 Oracle Coherence - DataGrid



title	Oracle Coherence - DataGrid
status	95
section	TBD
keywords	TBD

Oracle Coherence is an In-Memory Data Grid, data management platform for application objects that are shared across one or multiple distributed servers that requires low response time, very high throughput, predictable scalability, high availability and reliability [1196].

Initially this framework was developed by Tangosol. The Tangosol, was acquired by Oracle Corporation in 2007 [1197] and named that framework as Oracle Coherence.

Oracle coherence data grid platform used for computational intensive, stateful middle-tier applications that runs in a distributed platform. Coherence is directed to run application layer, and is often run in-memory with application itself. Oracle coherence Data Grid is a system of composed of multiple servers that work to manage information and related operations. Coherence provides the ideal infrastructure for building Data Grid services, and the client and server-based applications that use a Data Grid. At a basic level, Coherence can manage an huge amount of data across a large number of servers in a grid; it can provide minimal or low latency access for that data; it supports parallel queries to access the data with high throughput. Coherence also supports integration with database and EIS systems that act as the system of record for that data. Additionally, Coherence provides other services that are ideal for building more effective data grids [1196].

This framework comes by default with Oracle WebLogic 12c server.

Oracle also provides standalone coherence server, which can be used in any Big Data environment to store any database data, inflight or any dataset to be processed by MapReduce or any other Hadoop component. Cache servers can be configured in the clusters. This framework can be used in the Big Data environment without HDFS for in-memory distributed data storage. There are DOT NET, JAVA, C++ and other API's available along with REST service to access Coherence cache [1196]. Oracle coherence can be integrated from Spring based distributed applications as well [1198].

4.175 Oracle Nosql Database



title	Oracle Nosql Database
status	95
section	TBD
keywords	TBD

Oracle NoSQL Database[1199] is a scalable, distributed NoSQL database, designed to provide highly reliable, flexible and available data management across a configurable set of storage nodes. Data in Oracle NoSQL Database can be modeled as both relational-database-style tables, JSON documents or key-value pairs. Based on the hashed value of the primary key, Oracle NoSQL Database is a sharded s (shared-nothing) system which distributes the data uniformly across the multiple shards in the cluster. Storage nodes are replicated to ensure high availability, rapid failover in the event of a node failure and optimal load balancing of queries within each shard. NoSQL Database provides Java, C, Python and Node.js drivers and a REST API to simplify application development. A wide variety of related Oracle and open source applications are integrated in Oracle NoSQL Database, in order to simplify and streamline the development and deployment of modern big data applications. Oracle NoSQL Database is available in the following editions: Enterprise Edition - Oracle Commercial License Basic Edition - Oracle Database Enterprise Edition Commercial License Community Edition - Open source license.

4.176 Orange



title	Orange
status	95
section	TBD
keywords	TBD

Orange [1200] is a data mining, visualization, and machine learning toolkit based on Python 3. This software is developed in such a way as to allow practitioners to have varying degrees of technical background (including complete novices) and still utilize the product's capabilities [1200]. Orange has an interactive data visualization interface which allows for a simpler approach to perform complex data mining and machine learning practices and to derive insightful knowledge. Orange also provides a visual programming component in the form of widgets to perform qualitative analysis through a visualized workflow map. Depending on a widget's function, it is then grouped into a class, encouraging the use of various widgets in a typical given workflow. These widget visualizations also assist in the communication of analytic processes between domain experts and data scientists, which has encouraged the use of this product in academic and research settings [1200]; particularly with domains involving 'biomedicine, bioinformatics, [and] genomic research' [1201].

This open source toolkit's latest version (3+) uses various python libraries for computation, while utilizing the Qt framework for the visualization end [1201]. The available Python classes and methods include classes based on data models, preprocessing, classification, regression, clustering, distance, evaluation, and projection. The classification and regression classes offer the largest number of available methods, such as random forests, Naive Bayes, neural networks, and k-nearest neighbors [1200]. These classes can either be used directly as a Python library, or used in Orange's widget sets. It

is also possible to create custom widgets and include them in an Orange workflow [1201].

4.177 OrientDB hid-SP18-520



title	OrientDB
status	100
section	TBD
keywords	NoSQLDB, OrientDB

OrientDB is NoSQL database, which supports graph engines and it can be implemented as document database or as an Object-Oriented database [1202]. With Document database, it stores the documents as Key and value pairs. It refers group of these documents as collections. In OrientDB, classes or clusters are similar to Tables and every documents are similar to rows when compared with relational database models. With Graph databases, it creates the nodes known as Vertex and arcs known as Edges for storing the data. Vertex are equivalent to rows in relational database. Vertex and Edges properties are similar to the columns in relational database. It supports SQL as the built in query language. SB Tree [1203] is default indexing mechanism implemented with OrientDB for optimizing inserting of data and range queries. Reference and Embedded relationship is supported by OrientDB, where relationship with object is stored as direct links or the relationship is stored in the record itself[1204].

Graph databases are useful for developing application related to social networking and establish relationships between objects with respect to there properties. It maintains class relation using documents and links in document model [1205].

4.178 Owncloud



title	Owncloud
status	95
section	TBD
keywords	TBD

OwnCloud has made its significant impact in providing client-server software services for creating file hosting services and also to use them. Even though most of the functionalities are comparable to Dropbox, OwnCloud distinguishes itself by presenting as an open-source and free server edition. OwnCloud is easily available which makes any user easy to install and operate it.

OwnCloud is putting its best efforts to make it work like Google Drive, providing features such as online document editing, and contact synchronization [1206].

4.179 Paxata



title	Paxata
status	95
section	TBD
keywords	TBD

One of the most important and time consuming job of data scientist is to clean and prepare data from multiple sources in a format that it can be analyzed. Paxata [1207] semi automates the process by using its own algorithms. It uses machine learning and text mining combined with its libraries to efficiently clean data. Paxata provides a spreadsheet like interface where inconsistencies are color coded and instructions are provided to clean up data. Paxata visualizes the data in form of graphs and creates associations between various data objects and uses them to resolve data quality issues. This data can then be consumed by visualizing softwares like tableau. With this approach Paxata gives anyone ability to run data analytics on big data sets in a short amount of time.

this is also defined but not used in the text check: [1208]

4.180 Pig



title	Pig
status	95
section	TBD
keywords	TBD

Pig is a part of the Apache Hadoop ecosystem consisting of a scripting language called Pig Latin and a compiler that produces Map-Reduce programs. It was initially developed in 2006 at Yahoo! and taken over by Apache in 2007[1209]. Pig Latin allows developers to code multiple interrelated data transformations as data flow sequences, with the goal of making the code readable and easy to maintain. Pig optimized Pig Latin automatically and users can extend the language with purpose-written functions[1210]. There is some overlap in functionality between Pig and Hive, an SQL-like language that is also a part of the Apache Hadoop ecosystem. Pig tends to be favored by programmers and researchers, whereas Hive is preferred by data analysts[1211].

4.181 Pivotal



title	Pivotal
status	95
section	TBD
keywords	TBD

Pivotal is a developer of cloud-native applications, containers, and tools for DevOps. The primary cloud computing tool is the Pivotal Cloud Foundry (PCF) platform [1212]. PCF is a commercial platform built on the open-source Cloud Foundry platform. The architecture is container-based and offers an option to web developers in the shift to cloud-native software development [1213].

4.182 Pivotal Rabbit MQ



title	Pivotal Rabbit MQ
status	95
section	TBD
keywords	TBD

Pivotal RabbitMQ is a messaging broker platform used by various consumer applications like financial market data, system monitoring, business integration and various social, mobile, big data and cloud apps [1214]. Its protocol based nature lets it connect across various other software components making it an ideal messaging platform for cloud computing. It is efficient, scalable, portable across most operating systems [1214]. Its small disk and memory footprint makes it lightweight for use by developers. It has simple API and drivers are available for multiple languages like Python, PHP, Java. It also supports large scale messaging and routing according to topic and content [1214]. It is one of the new trending tools for many web applications.

4.183 Pool



title	Pool
status	95
section	TBD
keywords	TBD

Pool is a connection manager for R, which interfaces with the DBI family of connections [1015]. The advantage of using pool as a connection manager is that it automatically maintains a connection as open, or re-opens closed ones if needed. This helps ensure that for long-running, interactive contexts, such as data-visualization dashboards, access is maintained to data [1215]. Importantly, pool also closes connections at the end of session, ensuring that there are no dangling operations.

4.184 Apache PredictionIO



title	Apache PredictionIO
status	95
section	TBD
keywords	TBD

Apache PredictionIO is an open source machine learning stack for building, evaluating and deploying engines with machine learning algorithms. An open source Machine Learning Server built on top of an open source stack allows developers and data scientists to create predictive engines for any machine learning task. It allows developers to quickly build and deploy an engine as a web service and unify data from multiple platforms in batch or in real-time for comprehensive predictive analytics. It supports machine learning and data processing libraries such as Spark MLLib and OpenNLP [1216].

4.185 Presto



title	Presto
status	95
section	TBD
keywords	TBD

Presto is a SQL query engine developed specially for interactive analytics. It focuses on large commercial data warehouses with capacity of gigabytes to petabytes. It is open source and used for distributed systems. It is compatible with relational as well as NoSQL of data sources such as Cassandra and Hive [1217].

It is being used by big organizations like Facebook to run interactive queries against their large data warehouses. The main advantage of using Presto is that it allows to perform analytics on data from different data sources using single query. This allows data to be combined across organizations without extra overhead of separate queries for each data source [1217].

4.186 PubNub



title	PubNub
status	95
section	TBD
keywords	TBD

PubNub is globally recognized as a cloud Data Stream Network and a real-time infrastructure as a service platform founded by Stephen Blum and Todd Greene in 2010 [1218]. PubNub provides cloud-based services and products to build real-time web, mobile, and Internet of Things (IoT) applications [1219]. PubNub's main product is PubNub push messaging API which is currently being utilized by iOS, Android, Nodejs, and many other applications. This push messaging API is built on PubNub replicated global data streaming network at 14 data centers distributed among the entire world [1219]. PubNub is also being used as IoT device control platform to manage bidirectional communication, cross-device and platform messaging, monitor device metadata, act on data instantly, intelligent data routing, device provisioning and remote firmware upgrades, enterprise grade security, and minimal battery and bandwidth drain in home automation, wearables, connected car, sensor deployments, delivery and fulfillment, manufacturing and industrial, smart cities, and beacons and eTail [1218].

4.187 Pulsar



title	Pulsar
status	95
section	TBD
keywords	TBD

Apache Pulsar which is also an open source project of the Apache foundation was originally developed by Yahoo. It is a messaging solution that enables high performance server to server messaging. Similar to Kafka Pulsar is based on publisher-subscriber model. Some of the key features of Pulsar include low latency in publishing, guaranteed message delivery, scalability and so on. The publish-subscribe pattern involves components such as producers, consumers, topics and subscription wherein; topics are channels that transmit data from source to target or in other words from producers to consumers, producers job is to publish a message and a consumer process is the one that receives the message. Subscriptions are set of rules that determine how messages flow in the system from producers to consumers and have three modes namely exclusive, failover and shared [1220]. Pulsar can be installed and run in standalone mode or standalone cluster, it can also be run multiple clusters. Pulsar installation involves installing an instance which can be installed across clusters when installed in multi-cluster environment. In this setup clusters can be running within the data center or can span across multiple data centers. Pulsar also support geo-replication so the clusters can replicate with each other. Pulsar can also be installed on Kubernetes on Google Kubernetes or AWS [1220].

4.188 Puppet



title	Puppet
status	95
section	TBD
keywords	TBD

Puppet [684] is a open source software configuration and automation tool. It is written in C++ and Clojure. Puppet is a declarative language and uses domain specific language for configuration. Puppet uses facter to gather information about the system and user defines the desired state. Puppet does not use sequential programming where order of execution is key but uses graphical representation to represent the order of execution. Resources are defined in manifests written in Domain specific language. These manifests are complied into catalogue on puppet master and supplied to puppet clients. These catalogues are only applied if actual and desired states are different.

“Kubernetes [1221] is new cluster manager from google” and puppet makes it easy to manage the kubernetes resources.

Puppet is declarative, modular, has code testing features and therefore managing kubernetes with it is easier.

4.189 PyTorch



title	PyTorch
status	100
section	TBD
keywords	python, pytorch

PyTorch is known as Python based scientific computation package, built to support Deep learning researches and Neural network models. PyTorch has packages which are used for deep learning, multi processing, loading data for processing into the models. It is fast and has high computation speed when run with any size of datasets [1222].

PyTorch uses Tensors for building computation and are very similar to Numpy arrays. It uses the GPUs for its computation and responsible for faster retrieval of the data or results which are required for neural network. There are many variable provided by PyTorch which detail on Tensor and its gradient. It provides provision for computational graphs which helps a lot during the debugging of the code or viewing any results from any computations step by step [1223].

There are many modules defined in PyTorch, such as nn Module, which is acting as neural network layers producing outputs with weights details. Autograd module, which captures and helps calculating derivatives and gradients. With Optim module, which are very helpful when implementing the optimized model for neural network models creation [1224].

4.190 Qubole Data Service



title	Qubole Data Service
status	95
section	TBD
keywords	TBD

Qubole is a data services company who offers cloud services, data management, and data system security to companies. The BigData Platform in Qubole Data services contains all important elements for companies to process their data, and it is not hard to use. Multiple cloud providers can be chose, the company could select the cloud provider for their own commercial purpose (the most of the cloud providers in Qubole Data Service are open source) [1225].

The data analyze performance of the Qubole is the most competitive strength. In the Qubole, a company could uses the large scale cloud data platform, machine learning application, and other powerful tools to do data mining and analyzing. Even though the most of servers were cloud based, the claim from the Qubole is the security of the data and the system is guaranteed. Furthermore, the Qubole also provides different services (solutions) for different working type, different types of work could be done by different solutions, in order to match the biggest profit [1225].

4.191 RabbitMQ



title	RabbitMQ
status	100
section	TBD
keywords	messaging, Rabbitmq

RabbitMQ technology is open source message broker, which supports multiple messaging protocols. It has many features such as asynchronous messaging, which supports message queuing, receive and deliver acknowledgments, routing any message queues with broadcasting to logs or messages to multiple users [1226].

On the collection of nodes which is also known as clusters, Users shares the many resources among several nodes, where RabbitMQ applications are running and host them virtually, with managing queues and maintain runtime parameters [1227].

Several authentication mechanisms are been implemented and can be customized with RabbitMQ including SASL. By default, authentication in RabbitMQ is very PLAIN and requires setup on respective servers and clients [1228]. With the advancement of the technology which was earlier implemented as Advance Message Queuing and now it been enhanced for supporting streaming Text oriented messaging protocol. There are many other protocols have been developed to improve the technology [1229].

A browser UI based API is provided for monitoring and managing RabbitMQ servers. The admins can use this API for addressing and any kind of updates needed to provide better services [1230].

4.192 Rackspace



title	Rackspace
status	95
section	TBD
keywords	TBD

Rackspace is a cloud computing company which administers cloud system for their business partners and further helping them to concentrate on managing the business growth of their partners. Rackspace differentiates itself from other companies by stating that the cloud system alone which is provided by other companies are not sufficient to operate the infrastructure efficiently.

Rackspace proudly states that it takes several innovative and cognitive engineering skills to develop and manage the infrastructure and also concentrating on tools and applications which are necessary to administrate. BY providing all such toolwith updated data engines and e-commerce platforms, Rackspace claims that it will manage cloud and infrastructure of their business partners in a much different and efficient way [1231].

4.193 Ranger



title	Ranger
status	95
section	TBD
keywords	TBD

Apache Ranger [1232] is a framework to enable, monitor and manage comprehensive data security across the Hadoop platform. In order to provide comprehensive security across the Apache Hadoop ecosystem, the vision with Ranger was designed. The Hadoop platform can now support a true data lake architecture with the advent of Apache YARN. In a multitenant environment, Enterprises can potentially run multiple workloads. Data security within Hadoop needs to evolve to support multiple use cases for data access, while also providing a framework for central administration of security policies and monitoring of user access. Please read the FAQs if you need to understand how it works over Apache Hadoop components. The followings are goals of Apache Ranger: 1. Using a central UI or using REST APIs to centralize security administration to manage all security related tasks. 2. Use Hadoop components/tools to perform fine-grained authorization of specific operations and/or operations and manage them through central management tools. 3. Make sure that all Hadoop components authorization method are standardized. 4. Enhanced support for different authorization methods, such as role-based access control, attribute-based access control, etc. 5. Centrally audit user access and management operations (security-related) in all Hadoop components.

4.194 RapidMiner



title	RapidMiner
status	95
section	TBD
keywords	TBD

RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning and model deployment into production. It provides the facility to connect to various sources of different format and different scale to collect the data, data exploration to discover the pattern in the data, identify the issues, blending to find the relevant data for the modeling, clean the data for advanced algorithms. Once the data is prepared models are built using a nice workflow designer tool. It provides Classification, regression and clustering techniques. Validation is performed for cross validation and split validation after the model development for the accuracy of the model. On the successful validation of the model it is deployed into production with governance and security. It is Open and extensible and provide native R and Python support. The source code is available under an aGPL license. The platform provides an easy drag and drop environment for easy and fast data science related operations [1233].

4.195 Redis



title	Redis
status	100
section	TBD
keywords	redis, Redis, NoSQLDB

REmote DIctionary Server is No SQL database, supports Key value databases by mapping its key to type of values. Redis is an In-memory No SQL database, as it rely on the main memory to store data. It is an open source, and can support all kind of data structures similar to other databases including list, strings, ranges, bitmaps and many others. With in-memory data store feature, which helps Redis to provide a high performance and faster retrieval of the data. Redis can handle computation for one dimensional or multidimensional values[1234].

Redis has a feature for clearing the cache data once reached its max capacity defined during the configuration. It identifies the policies based on the different combination of trends and can be appropriately be removed[1235].

Redis has the feature named Redis replication which is very useful in case of connectivity issues between master and slaves servers, it provided the potential for making slaves as replica of master, in cases of any connectivity issues between master and slave servers or master server is not operational[1236].

4.196 RightScale Cloud Management



title	RightScale Cloud Management
status	95
section	TBD
keywords	TBD

RightScale Cloud Management is basically a platform which acts as a console to manage different clouds from one environment. Some of its features are automatic recovery protocols when it detects an escalation, disaster recovery architecture, automatic scaling and scripting.

“This platform facilitates ways to deploy and manage business-critical applications across public, private and hybrid clouds and provides configuration, monitoring, automation, and governance of cloud computing infrastructure and applications” [1237].

4.197 Ripple Transaction Protocol



title	Ripple Transaction Protocol
status	95
section	TBD
keywords	TBD

Ripple Transaction Protocol is an open-source protocol that allows transfer of anything of value (usually payments) on the internet.

"The Ripple network enables secure, instant and nearly free global financial transactions of any size with no chargebacks" [1238].

"Ripple provides one frictionless experience to send money globally using the power of blockchain" [1239].

There are 3 problems that ripple are trying to solve in the current payment systems available now high fees, charges for currency exchanges and processing delays.

Banks today built their own system for their customers that dont easily interact with each other. The Ripple network is designed to connect different payment systems together.

With ripple the transaction fees are very minimal and comparative less than what for example Visa charges. And there are no foreign exnchanges loses since the the currency is never converted to any other currency from the source to the destination. The transfer is also completed within minutes compared to the usual 3-business-days waiting period when sending money thru the traditional means.

4.198 Amazon SageMaker



title	Amazon SageMaker
status	95
section	TBD
keywords	TBD

Amazon SageMaker [1240] can help users be able to develop machine learning models in a more streamlined way. It includes functionality that allows a user to speed up the process of building, training and deploying their models. To help simplify the building of ML models, Jupyter notebooks are included which will allow the user to make it more convenient to explore and visualize training data stored in S3. The 12 most common machine learning algorithms have also come pre-installed and configured, as well as the frameworks Tensorflow and MXNet, with the option of also using your own specific framework. Using SageMaker to train your models allows you to scale up underlying infrastructure as needed up based on your storage needs. Automatic tuning of models is also included. SageMaker also takes advantage of EC2 in order to create highly available and elastic clusters where you can deploy your model. A/B testing capabilities are built into the product as well.

4.199 Sales Cloud



title	Sales Cloud
status	95
section	TBD
keywords	TBD

Sales Cloud is basically a part of the sales module of SalesForce. It is a platform which integrates the customer data together and it incorporates marketing, sales, customer service and business analytics functionalities. One of Sales Cloud's most important feature sets is

“sales performance management software. The sales performance management covers incentives, commissions, quotas, regions, goal setting, training and performance evaluation” [1241].

It also has features which enable us to construct dashboards and perform real time forecasting which are useful for data analytics. It has a mobile application of the same thereby providing more ease of access and portability.

4.200 Apache Samoa



title	Apache Samoa
status	95
section	TBD
keywords	TBD

Apache Samoa, which stands for Scalable Advanced Massive Online Analysis, is a distributed streaming machine learning framework that contains a programming abstraction for distributed streaming machine learning algorithms [1242].

“It features a pluggable architecture that allows it to run on several distributed stream processing engines such as Storm, S4, and Samza” [1242].

Real time analytics can be utilized by tools like Samoa and allow organizations to react in a timely manner when problems appear or to detect new trends helping to improve their performance by obtain useful knowledge from what is happening now [1243]. Apache Samoa users can develop distributed streaming ML algorithms once and execute them on multiple DSPEs (distributed stream processing engine) [1244]. In addition, users could also add new platforms by using the API provided, therefore, the Samoa project is divided into two different parts, namely: Samoa-API and Samoa-Platform. By using Samoa-API, developers could develop for Samoa without worrying about which DSPE is going to be used [1245]. Samoa, written in Java, is open source under the Apache Software License version 2.0.

4.201 Scikit-learn



title	Scikit-learn
status	95
section	TBD
keywords	TBD

A Google Summer of Code project, scikit-learn project is a machine-learning library. Written in Python, it is designed to be simple and efficient, accessible to non-experts, and reusable in various contexts. Its goal is to provide a set of common algorithms to Python users through a consistent interface.

This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language that is easy to follow.

“Emphasis is put on ease of use, performance, documentation, and API consistency. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings” [1246].

Although the library is easy to use its sophistication and power to analyze big data is never lost. This library has earned respect not only in the academe but also in the industry. There is increasing demand in the job market for scientists and engineers who specialize in this library. The author himself has personally used this library and can attest to its functionality. It has simplified a lot of usage of algorithms but it still is a powerful tool.

4.202 Scribe



title	Scribe
status	95
section	TBD
keywords	TBD

Scribe is a server design, originally developed and maintained by Facebook in 2008, that serves as an aggregation service for streaming log data. The Scribe server is deployed within each node of a network and sends the aggregated log data to a central server for analysis. The data is interpreted by the Scribe servers via a two-string input by a client: the category or direction, and the message itself. Scribe has been deployed on thousands of servers on a single network and is robust to network errors and failures [1247]. Scribe was developed by Facebook to prevent their highly distributed server architecture being locked into a third party's network topology. The purpose of Scribe is to solve two major needs of a distributed data system: capturing events, changes, and errors on the system, and maintaining the collected and aggregated data through issues common to decentralized networks such as connection breaks, server downtime, and scalability [1248]. The logging-functionality of Scribe is now maintained and improved upon through the open source community. Scribe is available via the Apache License v.2.

4.203 SETI @ Home



title	SETI @ Home
status	95
section	TBD
keywords	TBD

After funding was cut SETI lauched SETI @Home, a public volunteer computer via the internet. Using this software users donate idle CPU time for SETI to do calculations [1249]. It was released in 1999 and one of its goals was to prove the viability of volunteer computing. This goal has succeeded completly. SETI @Home was inspiration for several similar projects [1250], one of each is the LHC @home [1251].

4.204 ShareLatex



title	ShareLatex
status	95
section	TBD
keywords	TBD

ShareLatex is a cloud service accesable via a website. It allows real time collaboration and compilation of LaTeX documents [1252] as well as the storage of them.

ShareLatex has ease of use features, such as having several packages included in it server side. It provides several templates for presentations, papers among others. ShareLatex provides paid accounts as well as the free one. With the paid account you can see a history of the document and sinc the files to github or DropBox [1253].

4.205 Share Point



title	Share Point
status	95
section	TBD
keywords	TBD

Sharepoint is a web-based document management and storage system platform that integrates with Microsoft Office [1254]. It was launched in 2001 by Microsoft, now it has different editions with different functions.

Sharepoint allows for a few different applications. It can be used as a real time collaboration tool for Microsoft Office documents. Also providing a file history and keeping records. Sharepoint also provides a Social Network [1255], helping to centralize project management. It is integrated with Microsoft's OneDrive, allowing for mobility.

4.206 Skytap



title	Skytap
status	95
section	TBD
keywords	TBD

Skytap is a cloud platform that provides Environment-as-a-Service (EaaS). The company enables businesses to implement their IT without having to trouble themselves about the infrastructure needs of their products/services. One of the high lights of Skytap is that in addition to providing the cutting edge technologies in their environment, they cater to businesses that require traditional application or technologies. Beyond which they enable the customer businesses to modernize.

"True self-service, on-demand resources enable you to create your own software-defined datacenter and networks with environments on demand that work in the cloud just like in your datacenter." [1256]

4.207 Apache Solr



title	Apache Solr
status	95
section	TBD
keywords	TBD

Apache Solr is an open-source search platform used to build search applications. It leverages the power of Apache Lucene [1257], which is a Java-based search library providing the core operations required by any search application like Indexing and Searching [1258]. Apache Lucene and Apache Solr were merged together in 2010 and since then, produced by the Apache Software Foundation development team. It has an active development community and regular releases. Solr has RESTful API's like HTTP/XML [1104] or JSON [1105] to communicate with it that can be used from most popular programming languages [1259]. It has all capabilities required for a full-text search server such as tokens, spell check, wildcard, phrases and auto-complete. It is fast, highly scalable, reliable, fault-tolerant and enterprise-ready and can be deployed in any kind of systems such as standalone, distributed or cloud [1258]. Other major features include hit highlighting, built-in security, distributed search through sharding, database integration, faceted search, rich document (e.g., Word, PDF) handling [1259]. As Hadoop can handle large amounts of data, Solr can be used with Hadoop [1131] to find the required information from a large source. Apart from search, Solr also has NoSQL [1102] features and it can be used as a non-relational data storage and processing technology. The components of Solr can be customized easily by extending and configuring its Java classes thereby making it flexible and extensible [1258]. Solr provides navigation features to world's largest internet sites like Netflix, Instagram, Best Buy, eBay etc. [1260]. It is packaged as the built-in search in many applications such as content management systems and enterprise content management systems [1259].

4.208 SpagoBI



title	SpagoBI
status	95
section	TBD
keywords	TBD

SpagoBI is an open source business intelligence and big data analytics platform. The software is completely free, but paid user support, maintenance, consulting and training are available for purchase. It includes tools for [1261] reporting, multidimensional analysis (OLAP), charts, location intelligence, data mining, ETL and more. It also integrates with [1261] popular in-memory processing engines and enables real-time processing. SpagoBI allows analysis of [1262] unstructured data, such as as audio files, videos and images. It can also access different types of [1262] databases and analytical applications (such as Teradata), NoSQL databases (such as HBase) and HDFS (Hadoop) or distributions (Hortonworks)

4.209 Google Cloud Spanner



title	Google Cloud Spanner
status	100
section	TBD
keywords	TBD

Cloud Spanner is cloud based service for globally distributed database. It can grow horizontally and exponentially. Most data for business systems is stored in relational databases. There was need to bridge gap between relational and non relational database that provide high performance and availability. Cloud Spanner bridges this gap. This technology combines benefits of relational database structure such as atomicity, consistency, isolation, durability with non-relational databases that can scale horizontally. This is unique combination that allows transactions to be executed with high performance and greater consistency. Cloud Spanner revolutionizes database administration, management and makes application development more efficient. It is fully managed. It can be easily deployed and comes with out of the box synchronous and replication functionality[1263].

It takes advantages of all critical features of relational database such as schemas, ACID transactions, and SQL queries. It reduces need of high learning curve for developers who are well proficient in structured query language. Client libraries that can connect to spanner is language independent. These libraries can be developed in C sharp, Go, Java, Node.js, PHP, Python, and Ruby. Already existing JDBC driver with popular third-party tools can be used to connect with Google spanner[1263]. It is purposely built for global transactional consistency.

4.210 Spinnaker



title	Spinnaker
status	95
section	TBD
keywords	TBD

Spinnaker is an open source, multi-cloud continuous delivery platform that helps you release software updates with high velocity and confidence. It provides two core features: cluster management to view and manage your resources in the cloud and deployment management to construct and manage continuous delivery workflows [1264]. The main advantage of Spinnaker is it holds a modern software development concept of continuous delivery that is application updates should be delivered when they are ready, instead of on a fixed schedule. Also, it improves the speed, stability of application deployment processes along with supporting deployments across different platforms by several different cloud providers.

Although the project Spinnaker first started out with Netflix and then google joined in 2014, the Spinnaker community now includes dozens of organizations such as Microsoft, Oracle, Target, Veritas, Schibsted, Armory and Kenzan [1265].

4.211 SQLite



title	SQLite
status	95
section	TBD
keywords	TBD

SQLite is an open source, embedded relational database. Originally released in 2000, it has outstanding performance in terms of portability, ease of use, compactness, availability, and reliability [781].

SQLite has an exquisite, modular architecture and has introduced some unique methods for relational database management. It consists of eight independent modules organized in three subsystems [1266]. This model divides the query process into several discrete tasks, just like working on the assembly line. The query is compiled at the top of the architectural stack, executed in the middle, and the operating system's storage and interfaces are processed at the bottom.

4.212 Sqoop



title	Sqoop
status	95
section	TBD
keywords	TBD

The primary application of Sqoop is data transfer between the traditional or relational database management systems and Hadoop platforms. It also has the capability to transfer data from mainframes to Hadoop. Sqoop works with Oracle, MySQL and can import data from these sources into the Hadoop distributed File systems or HDFS. In addition it can also transform data in map-reduce or even export it to the database such as Oracle. Sqoop works in batch mode and cannot move data real time. [1267]. Sqoop relies on the database to describe the schema of the data being imported. It uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance. For databases, Sqoop reads the table row-by-row into HDFS. For mainframe datasets, Sqoop reads records from each mainframe dataset into HDFS. The output of this import process is a set of files containing a copy of the imported table or datasets. Since the import process runs in parallel processes each process creates a file causing multiple files being created. These text files can use different delimiters such as comma, pipe and so on [1267]. Sqoop relies on the database to describe the schema of the data being imported. It uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance. For databases, Sqoop reads the table row-by-row into HDFS. For mainframe datasets, Sqoop reads records from each mainframe dataset into HDFS. The output of this import process is a set of files containing a copy of the imported table or datasets. Since the import process runs in parallel processes each process creates a file causing multiple files being created. These text files can use different delimiters such as comma, pipe and so on [1267].

4.213 Stardog



title	Stardog
status	95
section	TBD
keywords	TBD

Stardog is a graph database from US-software company Complexible. Stardog has a particular focus on OWL and RDF-based systems, with the latest release Stardog 5.2 (9 January 2018) supports SPARQL query language; property graph model and Gremlin graph traversal language; OWL 2 and user-defined rules for inference and data analytics; virtual graphs; geospatial query answering; and programmatic interaction via several languages and network interfaces [1268]. Further, the developers of StarDog OWL/RDF DBMS have pioneered a new use of OWL as a schema language for RDF databases. This is achieved by adding integrity constraints (IC), also expressed in OWL syntax, to the traditional open-world OWL axioms [1269]. Other key features of Stardog include Machine Learning and Logical Inference, Semantic Search, Geospatial Search etc. As a commercial software, Stardog is priced for community, developer and enterprise tiers. The enterprise version has a free 30-day trial and the community version is free to download and use for up to four users and ten graph databases [1270].

4.214 Synthea



title	Synthea
status	95
section	TBD
keywords	TBD

Synthea [1271] is an open-source medical patient generator. Synthea allows for the full synthetic generation of medical patients and patient records, which solves the privacy problems of using real-world patient data. It also allows medical researchers to generate data on-demand and test scaling, stress, etc. Synthea uses a Generic Model Framework (GMF) to model and track disease progression as well. Each patient generated by Synthea is a full-model: from birth to present with full demographics. This type of data can be used for small and large-scale health analysis. The data underlying these models are generated based on current academic research. Therefore, the data can also be used to run analysis on the synthetic patients.

According to its website, Synthea is useful for Academic Research, the Health IT Industry, and Policy Formation. Synthea is a product of MITRE Corporation written in Java, and supports both C-CDA and FHIR formats. It can also generate graphs using Graphviz.

4.215 Synthetic Data Vault



title	Synthetic Data Vault
status	95
section	TBD
keywords	TBD

The most notable synthetic data generator is the Synthetic Data Vault (SDV)[1272]. Developed at MIT by Neha Patki, Roy Wedge, and Kalyan Veeramachaneni, the Synthetic Data Vault uses machine learning techniques to model database structure and content. The models can then be used to generate entirely synthetic tables and relationships which are true to the form of the originals. Because the synthetic data is generated and modeled mathematically according to the original data, very little, if any, insight is lost. Here we will explore why we should use synthetic data, how SDV generates synthetic data, and how to use the data generated by SDV.

SDV is written in Python and is, therefore, cross-platform. A separate file is required for each database in the table. Also, each database requires a configuration file in json format. The specifications for the configuration file will be shown later, but first we will discuss why such a product is necessary.

4.216 Apache SystemML



title	Apache SystemML
status	95
section	TBD
keywords	TBD

Apache SystemML is an [1273] open-source language and compiler that makes it dramatically easier to build custom machine learning solutions. Apache SystemML is [1273] flexible, scalable and optimal for Big Data that enables automatic optimization. SystemML's enables [1273] algorithm customizability via R-like and Python-like languages. It also has [1274] multiple execution modes, including Spark MLContext, Spark Batch, Hadoop Batch, Standalone, and JMLC. Its characteristics include [1274] automatic optimization based on data and cluster characteristics to ensure both efficiency and scalability.

4.217 Tableau



title	Tableau
status	95
section	TBD
keywords	TBD

Tableau is the data visualization software that helps people to see and understand the data. Tableau can connect to almost any database, drag and drop to create visualizations, and share with a click. We can either schedule to get data refreshed or have real time updates with live connection. We can explore data from any sources from spreadsheets to databases to Hadoop to cloud services in minutes and dashboard can be published live on the web and on mobile devices [1275]. Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. It also allows data blending and real-time collaboration, which makes it very unique. Tableau is also positioned as a leader Business Intelligence and Analytics Platform in Gartner Magic Quadrant.

4.218 Talend



title	Talend
status	95
section	TBD
keywords	TBD

Talend is an open source software that provides variety of tools for integration of data of an organization. It also helps to synchronize data between different systems. Some tools help to generate native code to deploy the data to hadoop. Talend is a drag on drop software helps to configure prebuilt components and clean the data from different sources. It also contains tools to check data quality for clients so that they can decide whether they need to clean the data before integrating it with clean datasets [1276].

4.219 TensorFlow



title	TensorFlow
status	100
section	TBD
keywords	Tensor, Tensorflow

TensorFlow provides a platform for implementing machine learning algorithms and is highly preferred with deep neural network models and algorithms. Google Brain team[hid-sp18-520-TensorFlowWiki] developed TensorFlow. TensorFlow is cross-platform, it can support mobile and embedded platforms and facilitates with APIs to support complex computations such as computation on the numeric data and generate the data flow graphs. TensorFlow has many powerful APIs which provides the ease with implementation of any model across any platforms, it can be implemented on a server, mobile or desktop with less configuration[1277]. TensorFlow provides high level API for training and build models such as Keras, Eager Execution, Estimators and Importing Data [1278].

Tensors represents the data in Tensorflow, which are multidimensional arrays. Model built using TensorFlow, can have lot of computations and steps before it produces the respective output for classification or prediction or any results for Deep learning algorithms and can be tedious to follow all the steps and calculations. TensorFlow provides a visualization of data flow of all these calculation and steps within TensorBoard suite. In dataflow graph, the nodes are mathematical operations and connecting edges provides the details on the data transferred from one node to other node[1279].

4.220 Teradata Intelliflex



title	Teradata Intelliflex
status	95
section	TBD
keywords	TBD

Teradata Intelliflex is an integrated environment for Data Warehouse functionalities which in its own way integrates some of the strategic and operational workload onto one Data warehouse. It is available on Intellicloud, which is a cloud offering of Teradata. Intelliflex can independently scale nodes enabling us to use nodes as required to manage the processing power and also

“store data on multiple layers of solid state drives”

with virtual storage as per our data requirements [1280].

4.221 Teradata Intellibase



title	Teradata Intellibase
status	95
section	TBD
keywords	TBD

Teradata Intellibase provides a compact environment to perform data warehousing, data exploration in an iterative way and advanced analytics using the stored data. Storage of data come at a low cost in Intellibase.

The platform enables a combination of Teradata and Hadoop nodes to make up for the varied workload requirements. It does this by installing everything into a single cabinet to preserve the floor space in the data center.

The features include Teradata Database, Hadoop, Teradata Aster Analytics and Teradata Unified Data Architecture. With these features, it enables application deployment in a single cabinet. It also provides advanced in-memory computing and also provides data protection for all the data sources. The hardware could be re-deployed elsewhere thereby reducing infrastructure costs. Additionally, it also provides software re-imaging for quick replication.

The technology specifications include 18 nodes, 375 TB of uncomprased data from user and 18TB of memory in a single cabinet. The processors are dual multi-core Intel Xeon Processors [1281].

4.222 Teradata Kylo



title	Teradata Kylo
status	95
section	TBD
keywords	TBD

Kylo is a data lake management software management platform. It is open source and provides features like data ingestion with data cleansing and validation, metadata management, governance and security.

It can connect to many data sources and infer the schema from the common available data formats. Kylo's data ingestion workflow transfers data from source to Hive tables with various configuration options which are built around validation of data fields, protection of data, data profiling, data security and overall governance.

Kylo includes a metadata repository and provides key capabilities for data exploration. Using this feature, users can search in data and metadata to explore their entities of interest to gain insights.

By utilizing Kylo's capabilities, designers can develop new pipeline templates in Apache Nifi. Kylo and Nifi can communicate between each other to handle tasks between the cluster and the data center. The combination of Kylo and Nifi enables data owners to create new data feeds [1282].

4.223 Theano



title	Theano
status	95
section	TBD
keywords	TBD

Theano is a numerical computation library built for Python programming language. Theano uses Numpy syntax for expressing computations that can be compiled to run efficiently on both CPU and GPU architectures. Theano is an open source project and most of the development is primarily contributed by a machine learning group at the University of Montreal [1283]. Theano is a Python library that allows developers to efficiently define, optimize, and evaluate mathematical expressions including multi-dimensional arrays. Some of the features of Theano include tight integration with NumPy, transparent use of a GPU to perform data intensive computations, dynamic C code generation for evaluating expressions faster and support for extensive unit testing and self-verification [1284].

4.224 The GO Programming Language



title	The GO Programming Language
status	95
section	TBD
keywords	TBD

Go is an open source programming language developed by team of Google Robert Griesemer, Rob Pike and Ken Thompson. Its easy to build simple, robust and efficient software. Go is clean and concise which directly compiles to machine code and has very efficient garbage collection mechanism.

It has been built to support concurrency mechanism for multiple processes to run inside a core efficiently. GOROUTINES, SELECT and CHANNELS are some primitive types like String in Java which are built around concurrent programming to solve complex problems.

Go is a cross platform and portable language which avoids several object oriented features like classes and inheritance which might not be suitable for multi-core processing and parallel computing.

Due to advent of big data technologies and cloud computing there had been a need of modern programming language which could address the efficiency and scalability requirements for these platforms. Go works really well in modern computing environment due to its really powerful and efficient dependency management built inside it which differentiates it from other languages like Java and C++.

Key features of GO Language are:

- Ease of construction
- High efficiency in working on large programs

- Many programmers can work on a code
- Concurrency
- Dependency management

[1285]

4.225 Tibco DataSynapse GridServer



title	Tibco DataSynapse GridServer
status	95
section	TBD
keywords	TBD

DataSynapse was founded by two ex-investment bankers with an idea to speed up calculations by running them in parallel, distributed over multiple machines in the cluster to avoid single point of failure. The first product live cluster was released in 2001. In 2004 this product was renamed to GridServer. Gridserver was developed to support larger and larger grid of network computers [1286]. DataSyapnse was acquired by Tibco in 2009 and later this product was renamed as: Tibco DataSynapse GridServer [1287].

DataSynapse GridServer is a highly scalable software infrastructure that allows application services to operate in a virtualized fashion, unattached to specific hardware resources. Client applications submit requests to the Grid environment and GridServer dynamically provisions services to respond to the request. Multiple client apps can submit multiple requests simultaneously. The GridServer dynamically creates multiple instances to handle requests in parallel on different Grid server nodes. This architecture is therefore highly scalable in both speed and throughput. A single client can see scalable performance gains in the processing of multiple requests, and many applications and users will see scalable throughput though there huge volumes of client requests [1288].

Data Synapse grid server has the capabilities of compute grid and data grid. The main components of the grid server are Engines, Directors and Brokers. All these components are JVM's built in Java. Each component has their own responsibilities. The applications are deployed in the engines and computation and processing is also done

in the engines. Engines are light weight containers. Directors receive the client requests and then navigate to the broker. Broker act as a load balancer to navigate the request to the available engines in the grid environment based on engine load, and availability. As there are multiple nodes in the grid, there will be primary and secondary director, broker, and several engines to support high availability and fault tolerance [1288].

4.226 TokuDB



title	TokuDB
status	95
section	TBD
keywords	TBD

TokuDB is an open-source storage engine for MySQL [1146] and MariaDB [1147] used for high-performance in write-intensive environments. It uses fractal-tree index data structure, that keeps the data sorted and allows searches and sequential data access simultaneously, thereby providing improved performance [1289]. TokuDB compresses all data on disk including indexes, thereby reducing the disk and flash-drive storage requirements. It eliminates slave lag with read free replication [1290]. It is ACID and MVCC compliant and offers online schema-modifications. It is also included in Percona server [1291] [1289]. The use of fractal-tree technology also enables TokuDB to speed indexing by 10 times or more, thereby improving the performance of large databases (typically 50 GB or more). Its exceptional indexing feature makes it an ideal solution for applications that must simultaneously query and update huge volumes of rapidly arriving data [1292]. This also makes it scalable and improves operational efficiency. TokuDB is well-suited for the demanding requirements of big data applications as it lowers the infrastructure costs associated with scaling and optimization efforts [1291]. It has zero-maintenance downtime which makes it highly available in both public and private environments including cloud [1290] [1291].

4.227 TreasureData



title	TreasureData
status	95
section	TBD
keywords	TBD

Treasure Data provide a platform to consolidate the customer data from different sources, integrate them and see actionable customer view. This is helpful for Marketing Analytics, Sales Operations Analytics. This helps in understanding customer behavior easily and fast and take the actions accordingly for business benefit. We need to access various level of historical and real time data to get 360 degree view of data. Treasure Data enables the connection to various sources easily and reduce the data cleaning process as it has its own in built tool for cleaning the data. Treasure Data work on structured and semi structured data. It has almost 100 connectors to connect to various sources including social media, pull the data in real time and maintain a single view of the customer using advanced machine learning technologies. Data is available within short time (in minutes) for the actions and business decision [1293].

4.228 Twilio



title	Twilio
status	95
section	TBD
keywords	TBD

Twilio is one among the famous companies which provide cloud communications platform as a services (PaaS). Through the APIs provided by Twilio, the software developers will be able to programmatically manage phone calls and also the text messages. It deploys its technology on the most successful HTTP protocol and also provides the flexibility of billing according to the usage.

In order to protect against unexpected outages, Twilio follows strict architectural design methodologies. For such efforts, Twilio has been applauded. Twilio also makes efforts in the development of open-source software and is consistently making contributions to the open-source community [1294].

4.229 US Consumer Financial Protection Bureau



title	US Consumer Financial Protection Bureau
status	95
section	TBD
keywords	TBD

The United States Consumer Financial Protection Bureau (CFPB) Data and Research organization is a consumer protection agency that was established after the 2008–2009 economic crisis. The agency was established to help enforce consumer protection laws and to help protect consumers against illegal financial risks. The agency has established multiple data sources that are free and open to the public for consumption and analysis [1295].

4.230 Google Vision



title	Google Vision
status	100
section	TBD
keywords	TBD

Google vision API help users quickly classify images into thousands of predefined meaningful categories. Image classification is complex and technically challenging task. It has been made simple by use of Vision API. Digital data such as images, voice, and video are stored and transmitted over network at much faster rate than ever before. Google Vision API provides platform that helps researchers and developers identify and categorize meaningful insights hidden in media formats with high performance and greater accuracy.

It does this by encapsulating powerful machine learning models such as KNN and Regression for classification of images. It help detects objects and faces within images. Vision API finds and reads printed words contained within images through Optical Character Recognition[1296].

Developers can define meta data for their image catalog. They can tag rating of content and define category such as adult content, offensive content etc using API. Marketing campaign may be launched through image sentiment analysis. It can be accessed through REST API. Document classification and product search are popular use cases of Cloud Vision other than image search. Label detection, Web detection, Logo detection, Handwriting recognition, Landmark detection, Face detection, and content moderation are most pupular features of Cloud Vision API [1296].

4.231 Weka



title	Weka
status	95
section	TBD
keywords	TBD

Weka [1297] is a machine learning environment that provides graphical interface. It has a library of machine learning algorithms for data mining tasks. It was developed in Waikato university New Zealand. Python and R have advanced tools for machine learning but can be difficult to learn specially for a beginner and Weka makes this transition easier. Weka is written in Java and has its own API. All tasks performed with GUI can be also performed in CLI@. Its GUI has Explorer to learn, experimenter to run experiments and Knowledge flow to build pipelines for actual implementations.

4.232 The World Bank



title	The World Bank
status	95
section	TBD
keywords	TBD

The world bank is a philanthropic organization that has two main goals. First, end extreme poverty and second, to promote shared prosperity. One of the ways the World Bank hopes to meet its goals is by developing and sharing an open-data platform that can be used by the public. The World Bank hopes the open data platform can be used to promote knowledge that will ultimately help with its goal of promoting prosperity and ending extreme poverty [1298].

4.233 WSO2 Analytics



title	WSO2 Analytics
status	95
section	TBD
keywords	TBD

WSO2 Analytics is provided by the WSO2 Stream Processor. The features of the stream processor are support for stream and event processing contracts, user friendly development interfaces, high availability and scalability, easy integration with other components and business friendly analytics dashboards [1299]. The stream processor is based on the Siddhi Processing engine [1300] and is capable of performing real-time analytics for different types of events. The query processing engine stays as the central component while the events pass through the engine. WSO2 drives the vision of digitizing businesses and analytics is a key part according to [1301]. The WSO2 team emphasizes the need for analytics in businesses where automation and analytics is the highlight while taking the maximum use of contextual data [1302].

4.234 XGBoost



title	XGBoost
status	95
section	TBD
keywords	TBD

“XGBoost is an open-source software library which provides the gradient boosting framework for C++, Java, Python, R, and Julia” [1303].

XGBoost stands for Extreme Gradient Boosting. Before talking about XGBoost, it is best to give introduction to general gradient boosting. Gradient Boosting is a machine learning technique used to build both regression and classification models. It is primarily used in building decision trees. But building gradient boosting models on huge datasets (that sometimes contain more than 500,000 observations) is computationally onerous, not so efficient.

“The name xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms. Which is the reason why many people use xgboost”.

- says Tianqi Chen, creator of XGBoost (later received contributions from many developers) [1304]. The description of XGBoost according to the software repository on github is

“Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more” [1305].

4.235 Zepplin



title	Zepplin
status	95
section	TBD
keywords	TBD

Apache [1306] Zepplin is open source web based notebook that has built in data discovery, exploration, visualization and collaborative features. Zeppelin interface is interactive and seamlessly provides a single interface to execute code and visualize in the same dashboard. Zepplin's architecture has three layers; frontend, Zepplin server and interpreter processor. Zeppelin's interpreter supports any language or data processing backend to act as input. It supports [1307] Apache Spark out of the box without any configuration.

4.236 Zmanda



title	Zmanda
status	95
section	TBD
keywords	TBD

Zmanda is an open source platform which offers open source cloud backup. The recovery software and the services provided by Zmanda are used by many of the small and the mid-size ventures. In order to effectively protect the Linux, Solaris, Windows, Mac OS X environments and enable the backup and recovery in these operating systems, Zmanda offers the Amanda Enterprise. The Zmanda Recovery Manager (ZRM) is targeted to achieve the functionalities for scheduling the full and incremental backups [1308].

There has been a huge growth in the data size in the recent years and numerous organizations lack the budget and don't have the ability to perform the complex tasks and manage the costly backups. In order to effectively address this, Zmanda provides the Amanda Enterprise which offers the backup and recovery services that integrates and provides the quick setup, disentangled administration to tasks, and less cost. Amand Enterprise liberates us from being bolted into a vendor by providing the standard formats and tools [1309].

Amanda Enterprise is one of the toll compelling and predominant commercial open source backup and recovery software. It provides the less time consuming solution with the goal to implement the backup tasks in a simplified manner for the various systems, databases and other applications. Apart, from these it also establishes the secure environment that puts a barrier for intruders to avoid breaching the critical data and the engineers can quickly restore the backups in a chaotic situation [1310].

References

- [1] G. von Laszewski, F. Wang, H. Lee, H. Chen, and G. C. Fox, "Accessing Multiple Clouds with Cloudmesh," in Proceedings of the 2014 acm international workshop on software-defined ecosystems, 2014, pp. 21–28 [Online]. Available: <http://doi.acm.org/10.1145/2609441.2609638>
- [2] Apache accumulo user manual version 1.8. [Online]. Available: https://accumulo.apache.org/1.8/accumulo_user_manual.html
- [3] ActiveBPEL, "Communicating with the activebpel server administration interface via web services." Web page, Feb-2017 [Online]. Available: http://www.activevos.com/content/developers/education/sample_acti
- [4] "ActiveMQ technology." Web page [Online]. Available: <http://activemq.apache.org/>
- [5] B. Synder and R. Davies, ActiveMQ in action, 1st ed. Manning publications, 2013.
- [6] Aerobatic, "Aerobatic - overview." Web page, Jan-2017 [Online]. Available: <https://www.aerobatic.com/docs/overview/>
- [7] L. Wang, P. V. Buren, and D. Ware, "Architecting a distributed bioinformatics platform with iRODS and iPlant agave api," in 2015 international conference on computational science and computational intelligence (csci), 2015.
- [8] "Agave api home - features tab." Web page [Online]. Available: <https://agaveapi.co/platform/features/>
- [9] A. S. Foundation, "Apache Airavata." Web page, 2016 [Online]. Available: <http://airavata.apache.org>
- [10] U. Project, "UltraScan Analysis Software." Web page, Apr-2015

[Online]. Available: <http://ultrascan.uthscsa.edu>

[11] Indiana University, Apache Airavata, XSEDE, and NSF, "SEAGrid Portal." Web page, Jul-2016 [Online]. Available: <https://seagrid.org>

[12] A. S. Foundation, "GenApp - Apache Airavata - Apache Software Foundation." Web page, Aug-2014 [Online]. Available: <https://cwiki.apache.org/confluence/display/AIRAVATA/GenApp>

[13] "Allegro." Web page [Online]. Available: <http://allegrograph.com/>

[14] "Allegrow." Web page [Online]. Available: <https://en.wikipedia.org/wiki/AllegroGraph>

[15] I. Amazon Web Services, "Amazon dynamodb." Web page [Online]. Available: <https://aws.amazon.com/dynamodb/>

[16] AWS, "Developer guide - limits in dynamodb." Web page [Online]. Available:

<http://docs.aws.amazon.com/amazondynamodb/latest/developerguid>

[17] "Kinesis - real-time streaming data in the aws cloud." Web page [Online]. Available: <https://aws.amazon.com/kinesis/>

[18] S. G. Shilpi Saxena, Real-time big data analytics, 1st ed. 35 Livery Street, Birmingham B3 2PB, UK: Packt Publishing, 2016.

[19] Amazon, "Amazon relational database service (rds)." Web page, Nov-2018 [Online]. Available: https://aws.amazon.com/rds/?nc1=h_ls

[20] J. Barr, "Introducing amazon rds - the amazon relational database service." Web page, Oct-2009 [Online]. Available: <https://aws.amazon.com/cn/blogs/aws/introducing-rds-the-amazon-relational-database-service/>

[21] T. Knaup, "MySQL in the cloud at airbnb." Web page, Nov-2010 [Online]. Available: <https://medium.com/airbnb-engineering/mysql-in-the-cloud-at-airbnb-336e5666bc94>

- [22] Amazon, "What is amazon relational database service (amazon rds)?" Web page, Oct-2014 [Online]. Available: <https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html>
- [23] R. Thelwell, "What is amazon redshift? AWS's "fastest growing service" unboxed." Online [Online]. Available: <https://www.matillion.com/blog/redshift/what-is-amazon-redshift-aws/>
- [24] A. W. Services, "Introduction to amazon redshift - data warehouse solution on aws." Youtube, Feb-2015 [Online]. Available: <https://www.youtube.com/watch?v=AUvn49gey8Y>
- [25] A. Brust, "Amazon redshift: ParAccel in, costly appliances out." Online, Nov-2012 [Online]. Available: <https://www.zdnet.com/article/amazon-redshift-paraccel-in-costly-appliances-out/>
- [26] "Amazon route 53." Web page [Online]. Available: https://en.wikipedia.org/wiki/Amazon_Route_53
- [27] "What is amazon route 53?" Web page [Online]. Available: <http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/Welcome.html>
- [28] "Amazon s3." Web page [Online]. Available: <https://aws.amazon.com/s3/>
- [29] "Using Amazon S3." Web page [Online]. Available: <http://docs.aws.amazon.com/AmazonS3/latest/gsg/CopyingAnObject.html>
- [30] Amazon, "Amazon simple notification service." Web page [Online]. Available: <https://aws.amazon.com/sns/>
- [31] Amazon, Inc., "Amazon web services." Web page, Oct-2018 [Online]. Available: <https://aws.amazon.com>
- [32] Hortonworks, "Apache ambari overview." [Online]. Available: <https://hortonworks.com/apache/ambari/>

- [33] S. M, "01. Hadoop administration tutorial - ambari overview." Jul-2016 [Online]. Available: <https://www.youtube.com/watch?v=VDAh-YzUMm4&t=249s>
- [34] CloudClump, "Introduction to ambari." Oct-2016 [Online]. Available: <https://www.youtube.com/watch?v=W3Vqz06Djtw>
- [35] IntelliPaat, "Introduction to apache ambari." Jun-2018 [Online]. Available: <https://intellipaat.com/blog/what-is-apache-ambari/>
- [36] J. O'Hara, "Amqp." 2018 [Online]. Available: <https://www.techopedia.com/definition/26456/advanced-message-queuing-protocol-amqp>
- [37] J. O'Hara, "Amqp." 2018 [Online]. Available: <https://www.digitalocean.com/community/tutorials/an-advanced-message-queuing-protocol-amqp-walkthrough>
- [38] Red Hat, Inc., "Ansible documentation." Web page, Feb-2015 [Online]. Available: <https://docs.ansible.com/ansible/index.html>
- [39] Wikipedia, "Ansible (software)." Web page, Feb-2017 [Online]. Available: [https://en.wikipedia.org/wiki/Ansible_\(software\)](https://en.wikipedia.org/wiki/Ansible_(software))
- [40] J. Wettinger, U. Breitenbücher, and F. Leymann, "Any2API - automated apification," in CLOSER 2015 - proceedings of the 5th international conference on cloud computing and services science, 2015, pp. 475–486 [Online]. Available: <https://pdfs.semanticscholar.org/1cd4/4b87be8cf68ea5c4c642d38678>
- [41] J. Wettinger, "any2api - the better way to create awesome apis." Web page [Online]. Available: <http://www.any2api.org/>
- [42] "The Apache Software Foundation." Web page [Online]. Available: <http://ant.apache.org/>
- [43] Wikipedia, "Apache apex wiki." Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_Apex

- [44] Apache Apex, "Operator development guide." Web page [Online]. Available: https://apex.apache.org/docs/apex/operator_development/
- [45] D. Chan, "Big data with apache apex." Web page, Jan-2016 [Online]. Available: <https://jaxenter.com/big-data-apache-apex-122839.html>
- [46] Apache, "Apache apex malhar." Web page [Online]. Available: <http://apex.apache.org/docs/malhar/>
- [47] A. Kekre, "Apache apex blog." Web page, Sep-2015 [Online]. Available: <https://www.datatorrent.com/blog/introducing-apache-apex-incubating/>
- [48] Apache, "Apache arrow." Web page [Online]. Available: <http://arrow.apache.org/>
- [49] W. Mckinney, "Apache arrow: Cross-language development platform for in-memory data." Conference SciPy 2018, 2018 [Online]. Available: <https://www.slideshare.net/wesm/apache-arrow-crosslanguage-development-platform-for-inmemory-data-105427919>
- [50] M. Kornacker, T. Lipcon, and W. Mckinney, "Introducing apache arrow: A fast, interoperable in-memory columnar data structure standard." cloudera, Feb-2016 [Online]. Available: <http://blog.cloudera.com/blog/2016/02/introducing-apache-arrow-a-fast-interoperable-in-memory-columnar-data-structure-standard/>
- [51] S. Kaif, "What is apache beam?" Oct-2018 [Online]. Available: <https://www.quora.com/What-is-Apache-Beam>
- [52] T. A. S. Fondation, "Apache beam overview." [Online]. Available: <https://beam.apache.org/get-started/beam-overview/>
- [53] A. Woodie, "Apache beam's ambitious goal: Unify big data development." Web page, Apr-2016 [Online]. Available: <https://www.datanami.com/2016/04/22/apache-beam-emerges-ambitious-goal-unify-big-data-development/>

[54] Apache, “Apache derby.” Web Page, 04-Nov-2018 [Online]. Available: <https://db.apache.org/derby/>

[55] Wikipedia, “Apache derby.” Web Page, 18-Oct-2018 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Derby

[56] Microsoft, “MS sql server.” Web Page, 04-Nov-2018 [Online]. Available: <https://www.microsoft.com/en-us/sql-server/sql-server-2017>

[57] Oracle, “Oracle database.” Web Page, 04-Nov-2018 [Online]. Available:

<https://www.oracle.com/technetwork/database/windows/index-088762.html>

[58] P. C. Zikopoulos, G. Baklarz, and D. Scott, “Apache derby – off to the races: Includes details of ibm cloudscape,” Pearson, 2005, pp. 8-11 [Online]. Available: <http://www.informit.com/articles/article.aspx?p=422309&seqNum=8>

[59] A. S. Foundation, “About apache flex.” Web page, Feb-2017 [Online]. Available: <http://flex.apache.org/about-whatis.html>

[60] J. Jackson, “Adobe donates flex to apache,” The IDG News Service, Nov. 2011 [Online]. Available: https://www.techworld.com.au/article/407714/adobe_donates_flex_ap

[61] Wikipedia, “Apache flex.” Web page, Mar-2017 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Flex

[62] Web page [Online]. Available: <http://hawq.apache.org/>

[63] Pivotal, “Pivotal life cycle matrix.” Web page, 2018 [Online]. Available: <https://d1fto35gcfffzn.cloudfront.net/support/PivotalLifecycleMatrix.pdf>

[64] Apache Knox, “Apache knox.” Web page, Feb-2017 [Online]. Available: <https://knox.apache.org/>

- [65] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Doctoral Dissertation, University of California, Irvine, 2000 [Online]. Available: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
- [66] S. Peyrott, "API gateway. An introduction to microservices, part 2." Web page, Sep-2015 [Online]. Available: <https://auth0.com/blog/an-introduction-to-microservices-part-2-API-gateway/>
- [67] "Apache OODT." Web page [Online]. Available: <http://oodt.apache.org/>
- [68] C. Mattmann, "A look into the apache oodt ecosystem." Nov-2011 [Online]. Available: <https://www.slideshare.net/chrismattmann/a-look-into-the-apache-oodt-ecosystem>
- [69] J. Fan, J. Yan, Y. Ma, and L. Wang, "Big data integration in remote sensing across a distributed metadata-based spatial infrastructure," *Remote Sensing*, vol. 10, no. 2, p. 7, Dec. 2017.
- [70] T. A. S. Foundation, "Apache oodt." Feb-2010 [Online]. Available: <https://cwiki.apache.org/confluence/display/OODT/Home>
- [71] Apache Software Foundation, "Apache ranger." Web page [Online]. Available: <http://ranger.apache.org/>
- [72] Hortonworks, "Apache ranger - overview." Web page [Online]. Available: https://hortonworks.com/apache/ranger/#section_2
- [73] S. Mahmood and S. Venkat, "FOR your eyes only: DYNAMIC column masking & row-level filtering in hdp2.5." Web page, Sep-2016 [Online]. Available: <https://hortonworks.com/blog/eyes-dynamic-column-masking-row-level-filtering-hdp2-5/>
- [74] A. org., "Apache tomcat." Web page [Online]. Available: <http://tomcat.apache.org/>
- [75] Wikipedia, "Apache tomcat." 2010 [Online]. Available:

https://en.wikipedia.org/wiki/Apache_Tomcat

[76] P. K. Wee, "Instant appfog," Packt Publishing, 2013 [Online]. Available: <https://www.packtpub.com/mapt/book/Web-Development/9781782167624/1/ch01lvl1sec03/So,+what+is+AppFog%>

[77] S. B. Margaret Rouse, "Platform as a service (paas)." Web page, Sep-2017 [Online]. Available: <https://searchcloudcomputing.techtarget.com/definition/Platform-as-a-Service-PaaS>

[78] S. B. Margaret Rouse, "Platform as a service (paas)." Web page, Sep-2017 [Online]. Available: <https://searchcloudcomputing.techtarget.com/definition/Platform-as-a-Service-PaaS>

[79] CenturyLink, "CenturyLink appfog." Web page, 2016 [Online]. Available: http://www.centurylink.com/asset/business/enterprise/reference-architecture/ra-centurylink-paas_appfog_cm160194.pdf

[80] techcrunch, "CenturyLink acquiring appfog to move into platform-as-a-service market." Web page [Online]. Available: <https://techcrunch.com/2013/06/13/centurylink-acquiring-appfog-to-move-into-platform-as-a-service-market/>

[81] L. Carlson, Programming for paas: A practical guide to coding for platform-as-a-service. O'Reilly, 2014 [Online]. Available: https://books.google.com/books?id=7Z1_AAAAQBAJ&pg=PP2&dq=centurylink+appfog&hl=en&sa=X&ved=0ahUKEwzvqNjBxLXRAhVgj4KHSWzDyAQ6wEw

[82] A. Abiola, "CenturyLink: Can it show growth with the appfog acquisition?" Web page, Jan-2013 [Online]. Available: <https://seekingalpha.com/article/1507842-centurylink-can-it-show-growth-with-the-appfog-acquisition>

[83] AppScale, "What-is-appscale." Web page, 2016 [Online]. Available: <https://www.appspot.com/community/what-is-appscale/>

[84] Appscale, "Why-use-appscale." Web page, 2016 [Online]. Available: <https://www.appscale.com/get-started/deployment-types/>

[85] F. Pop, J. Kolodziej, and B. DiMartino, "Resource management for big data platforms," Springer Nature, 2016, pp. 49–50.

[86] "At1." Web page [Online]. Available: <http://www.cyverse.org/atmosphere>

[87] Wikipedia, "Virtual machine." Web page [Online]. Available: https://en.wikipedia.org/wiki/Virtual_machine

[88] Cyverse, "Atmosphere." Web page [Online]. Available: <https://wiki.cyverse.org/wiki/display/atmman/About+Atmosphere>

[89] "Apache Avro." Web page [Online]. Available: <https://avro.apache.org/docs/current/>

[90] C. Toh, "AWS elastic beanstalk survival guide." Web, Jan-2016.

[91] J. Nutt, "How to deploy a node.js app to the aws elastic beanstalk," freeCodeCamp, 2018.

[92] M. Rouse, "AWS elastic beanstalk," SearchAWS, 2014.

[93] Wikipedia, "Chef (software)." Web page, Jan-2017 [Online]. Available: [https://en.wikipedia.org/wiki/Chef_\(software\)](https://en.wikipedia.org/wiki/Chef_(software))

[94] Amazon, "AWS opsworks." Web page [Online]. Available: <https://aws.amazon.com/opsworks/>

[95] "An Introduction to Windows Azure BLOB Storage," Simple Talk. Web page, 13-Feb-2017 [Online]. Available: <https://www.simple-talk.com/cloud/cloud-data/an-introduction-to-windows-azure-blob-storage/>

[96] "Get started with Azure Blob storage (object storage) using .NET Microsoft Docs." Web page, 13-Feb-2017 [Online]. Available: <https://docs.microsoft.com/en-us/azure/storage/storage-dotnet-how->

to-use-blobs

- [97] S. Lo, D. Laudenschlager, C. Casey, S. Pelluru, and A. Narain, "Introduction to azure data factory." Web page, Jan-2018 [Online]. Available: <https://docs.microsoft.com/en-us/azure/data-factory/introduction>
- [98] Microsoft, "Data factory product overview." Web page [Online]. Available: <https://azure.microsoft.com/en-us/services/data-factory/>
- [99] J. Hudiono, "How to choose the right etl tool for your business." Web page, Aug-2017 [Online]. Available: <https://blog.chartio.com/posts/how-to-choose-the-right-etl-tool-for-your-business>
- [100] MuleSoft, "MuleSoft anypoint exchange." Web page [Online]. Available: <https://www.mulesoft.com/exchange/>
- [101] Amazon, "AWS glue pricing." Web page [Online]. Available: <https://aws.amazon.com/glue/pricing/>
- [102] Microsoft, "Microsoft data factory pricing." Web page [Online]. Available: <https://azure.microsoft.com/en-us/pricing/details/data-factory/data-pipeline/>
- [103] C. Gronlund, G. Ericson, L. Francs, and P. McKay, "Azure machine learning." Web page, Jan-2017 [Online]. Available: <https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-what-is-machine-learning#what-is-machine-learning-in-the-microsoft-azure-cloud>
- [104] A. Silberschatz, P. B. Galvin, G. Gagne, and A. Silberschatz, Operating system concepts, vol. 4. Addison-wesley Reading, 1998.
- [105] "Get started with azure queue storage using .NET." Web page [Online]. Available: <https://docs.microsoft.com/en-us/azure/storage/storage-dotnet-how-to-use-queues>
- [106] "Microsoft azure - queues." Web page [Online]. Available:

https://www.tutorialspoint.com/microsoft_azure/microsoft_azure_que

[107] M. Rouse, “SQL azure.” Web page, 2011 [Online]. Available: <https://searchsqlserver.techtarget.com/definition/SQL-Azure>

[108] C. Rabeler, “The azure sql database service.” 2018 [Online]. Available: <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-technical-overview>

[109] B. LUIJBREGTS, “Compare azure sql database vs. Azure sql data warehouse: Definitions, differences and when to use.” Web page, 2018 [Online]. Available: <https://stackify.com/azure-sql-database-vs-warehouse>

[110] R. Sheldon, “Take bi up a notch with sql azure reporting.” 2011 [Online]. Available: <https://searchsqlserver.techtarget.com/feature/Take-BI-up-a-notch-with-SQL-Azure-Reporting>

[111] “Microsoft Azure real-time data analytics.” Web page [Online]. Available: <https://azure.microsoft.com/en-us/services/stream-analytics/>

[112] “Microsoft Docs azure stream analytics documentation.” Web page [Online]. Available: <https://docs.microsoft.com/en-us/azure/stream-analytics/>

[113] “Github azure/ azure-stream-analytics.” Web page [Online]. Available: <https://github.com/Azure/azure-stream-analytics/>

[114] Microsoft Corp., “Azure.” Web page, Jan-2017 [Online]. Available: <https://azure.microsoft.com/en-us/>

[115] J. Hassell, “Microsoft azure wikipedia.” Web page, 2014 [Online]. Available: https://en.wikipedia.org/wiki/Microsoft_Azure

[116] Wikipedia, “Cloud computing.” Web page, Jan-2017 [Online]. Available: https://en.wikipedia.org/wiki/Cloud_computing

- [117] Microsoft Corp., "Form 10-k." Web page, Jul-2016 [Online]. Available:
<https://www.sec.gov/Archives/edgar/data/789019/0001193125166622>
- [118] Amazon.com, Inc., "Amazon web services." Web page, Jan-2017 [Online]. Available: <https://aws.amazon.com>
- [119] "Oracle website." Web page [Online]. Available: <http://www.oracle.com/technetwork/database/database-technologies/berkeleydb>
- [120] "What is berkeley db?" Web.
- [121] K. B. Margo Seltzer, "Berkeley db," The Architecture of Open Source Applications: Elegance, Evolution, and a Few Fearless Hacks, 2010.
- [122] R. C. Gentleman et al., "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004 [Online]. Available: <http://dx.doi.org/10.1186/gb-2004-5-10-r80>
- [123] "About bioconductor." Web page [Online]. Available: <https://www.bioconductor.org/about/>
- [124] bioKepler, "What is bioKepler." Web page [Online]. Available: <http://www.biokepler.org/faq#what-is-biokepler>
- [125] bioKepler, "Demo workflow." Web page [Online]. Available: <http://www.biokepler.org/userguide#demos>
- [126] W. Li, "Introduction to bioActors," in BioKepler tools and its applications, 2012, p. 31 [Online]. Available: <http://www.biokepler.org/sites/swat.sdsc.edu.biokepler/files/workshop09-05/slides/2012-09-05-02-Li.pdf>
- [127] "BITTORRENT." Web page, 2017 [Online]. Available: <https://www.lifewire.com/how-torrent-downloading-works-2483513>

- [128] Continuum Analytics, “Blaze.” Web page, 2015 [Online]. Available: <http://blaze.readthedocs.io/en/latest/index.html#>
- [129] Blaze, “The blaze ecosystem.” Web page, 2016 [Online]. Available: <http://blaze.pydata.org>
- [130] P. Howard, “Blazegraph gpu,” 2015, p. 13 [Online]. Available: https://www.blazegraph.com/whitepapers/Blazegraph-gpu_InDetail_BloorResearch.pdf
- [131] Systap, “Blazegraph wiki.” Web page, Aug-2008 [Online]. Available: https://wiki.blazegraph.com/wiki/index.php/Main_Page
- [132] BlinkDB, “BlinkDB.” [Online]. Available: <http://blinkdb.org>
- [133] S. Agarwal, A. Panda, B. Mozafari, S. Madden, and I. Stoica, “BlinkDB: Queries with bounded errors and bounded response times on very large data,” CoRR, vol. abs/1203.5485, 2012 [Online]. Available: <http://arxiv.org/abs/1203.5485>
- [134] IBM, “Exercise: Analyze business processes with ibm bpm blueprint.” Web page [Online]. Available: <http://www.ibm.com/developerworks/downloads/soasandbox/bluepri>
- [135] IBM, “BlueworksLive.” Web page [Online]. Available: <https://www.blueworkslive.com/home>
- [136] IBM, “IBM blueworks live.” Web page [Online]. Available: https://en.wikipedia.org/wiki/IBM_Blueworks_Live
- [137] M. Garnaat, “boto components.” Web page [Online]. Available: <http://boto.cloudhackers.com/en/latest/>
- [138] M. Garnaat, “boto-github components.” Web page [Online]. Available: <https://github.com/boto/boto3>
- [139] M. Garnaat, “boto-amazon-python-sdk components.” Web page [Online]. Available: <https://aws.amazon.com/sdk-for-python/>

- [140] M. Garnaat, "boto3-documentation components." Web page [Online]. Available: <https://boto3.readthedocs.io/en/latest/>
- [141] M. Plassnig, "Heroku-style application deployments with docker - dzone cloud." Web page, Nov-2015 [Online]. Available: <https://dzone.com/articles/heroku-style-application-deployments-with-docker>
- [142] J. Gonzalez and J. Lindsay, "Buildstep." Code repository, Jul-2015 [Online]. Available: <https://github.com/progium/buildstep>
- [143] S. Yangqing Jia Evan, "Caffe | deep learning framework." Web page [Online]. Available: <http://caffe.berkeleyvision.org/>
- [144] "Cascading." Web Page [Online]. Available: <https://www.cascading.org/>
- [145] "Apache cassandra." Web page, 2016 [Online]. Available: <http://cassandra.apache.org/>
- [146] CASK, "CASK - the first unified integration platform for big data." Web page [Online]. Available: <http://cask.co/products/cdap/>
- [147] CASK, "Getting started developing with cdap." Web page [Online]. Available: <http://docs.cask.co/cdap/current/en/developers-manual/getting-started/index.html>
- [148] "CDAP applications." Code Repository, May-2015 [Online]. Available: <https://github.com/caskdata/cdap-apps>
- [149] NASA/GSFC, "CDF home page." NASA,Goddard Space Flight Center; Web page, 2017 [Online]. Available: <http://cdf.gsfc.nasa.gov/>
- [150] NASA/GSFC, CDF user's guide (v3.3.6), Version 3.3.6. NASA / Goddard Space Flight Center, Greenbelt, Maryland 20771 (U.S.A.): NASA/GSFC Space Physics Data Facility, 2016 [Online]. Available: <http://spdf.gsfc.nasa.gov/pub/software/cdf/doc/cdf363/cdf363ug.pdf>
- [151] "CDF - common data format (multidimensional datasets)." Web

- page, Mar-2014 [Online]. Available:
<http://www.digitalpreservation.gov/formats/fdd/fdd000226.shtml#use>
- [152] Wikipedia, “Hierarchical data format.” Web page, Feb-2017 [Online]. Available:
https://en.wikipedia.org/wiki/Hierarchical_Data_Format
- [153] S. N. I. Association, “Cloud data management interface.” Web page; snia.org, Mar-2015 [Online]. Available:
<https://www.snia.org/cdmi>
- [154] Wikipedia, “Cloud data management interface.” 2018 [Online]. Available:
https://en.wikipedia.org/wiki/Cloud_Data_Management_Interface
- [155] S. K, “Cloud data management interface (cdmi) media types.” Sep-2010 [Online]. Available: <https://tools.ietf.org/id/draft-cdmi-mediatypes-02.xml>
- [156] M. Rouse, “Cloud data management interface.” Apr-2012 [Online]. Available:
<https://searchstorage.techtarget.com/definition/Cloud-Data-Management-Interface>
- [157] “Celery.” Web page [Online]. Available:
<http://www.celeryproject.org/>
- [158] “Celery - distributed task queue.” Web page [Online]. Available:
<http://docs.celeryproject.org/en/latest/index.html>
- [159] Ceph, “Ceph architecture.” Web page, Oct-2018 [Online]. Available: <http://docs.ceph.com/docs/master/architecture/>
- [160] Ceph, “Ceph block storage.” Web page, Oct-2018 [Online]. Available: <https://ceph.com/ceph-storage/block-storage/>
- [161] Ceph, “Intro to ceph.” Web page, Oct-2018 [Online]. Available:
<http://docs.ceph.com/docs/master/start/intro/>

[162] Tutorialspoint, "Chef quick guide." Web page, Oct-2018 [Online]. Available: https://www.tutorialspoint.com/chef/chef_quick_guide.htm

[163] "Cinder - openstack." Web page [Online]. Available: <https://wiki.openstack.org/wiki/Cinder>

[164] D. Radez, OpenStack essentials. Packt Publishing Ltd., 2015 [Online]. Available: <http://ebook.konfigurasi.net/Openstack/OpenStack%20Essentials.pdf>

[165] "CINET - cyberinfrastructure for network science." Web page [Online]. Available: www.bi.vt.edu

[166] Keith-Bisset, "CINET - a cyber-infrastructure for network science." Web page [Online]. Available: www.portal.futuresystems.org/project/233

[167] "Cloud and systems management." Web page [Online]. Available: <http://www.cisco.com/c/en/us/products/cloud-systems-management>

[168] S. Miniman, "Cisco moves up the cloud stack with intelligent automation." Web page, 2011 [Online]. Available: http://wikibon.org/wiki/v/Cisco_Moves_Up_the_Cloud_Stack_with_Intel

[169] W. Contributors, "Cloud foundry." Web page, Oct-2018 [Online]. Available: https://en.wikipedia.org/wiki/Cloud_Foundry

[170] D. C. E. Winn, "Chapter 1. Introduction," in Cloud foundry-the cloud native platform, O'Reilley Media, 2016 [Online]. Available: <https://www.oreilly.com/library/view/cloud-foundry/9781491965771/ch01.html>

[171] C. F. Foundation, "Cloud foundry concepts." Web page, 2018 [Online]. Available: <https://docs.cloudfoundry.org/concepts/>

[172] V. BADOLA, "What is cloud foundry? Key benefits and a real use case." Web page, Sep-2015 [Online]. Available: <https://cloudacademy.com/blog/cloud-foundry-benefits/>

- [173] “Cloudbees wikipedia documentation.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/CloudBees>
- [174] “Cloudbees webpage documentation.” Web page [Online]. Available: <https://www.cloudbees.com/products>
- [175] Cloudmesh.org, “HPC.” Web page [Online]. Available: <http://cloudmesh.github.io/hpc.html>
- [176] Cloudmesh.org, “Rain.” Web page [Online]. Available: <http://cloudmesh.github.io/rain.html>
- [177] G. von Laszewski, “Cloudmesh overview.” Web page, 2015 [Online]. Available: http://cloudmesh.github.io/introduction_to_cloud_computing/cloudmesh_overview.html
- [178] SINTEF, “CloudML.” Web page, 2013 [Online]. Available: <http://cloudml.org/>
- [179] SINTEF Git, “CloudML wiki.” Web page [Online]. Available: <https://github.com/SINTEF-9012/cloudml/wiki>
- [180] A. S. Foundation, “Cloudstack concepts and terminology.” Web page, 2018 [Online]. Available: <http://docs.cloudstack.apache.org/en/4.11.1.0/conceptsandterminology.html>
- [181] V. Ramesh, “Apache cloudstack tutorial : Key feature & architecture overview - part 3.” Video, Oct-2017 [Online]. Available: https://www.youtube.com/watch?v=SkE54b5q0J0&list=PLonJJ3BVjZW7Uqg7yjcWc_0rg1rDfZAgH&index=2
- [182] A. S. Foundation, “Cloudstack’s history.” Web page, 2018 [Online]. Available: <https://cloudstack.apache.org/history.html>
- [183] A. S. Foundation, “Apache cloudstack users.” Web page, 2017 [Online]. Available: <https://cloudstack.apache.org/users.html>
- [184] “CNTK/CNTKBook.” Web page [Online]. Available: <https://github.com/Microsoft/CNTK/blob/master/Documentation/CNTKBook.md>

<TechReport/lyx/CNTKBook-20160217.pdf>. [Accessed: 13-Feb-2017]

[185] "Home - Microsoft/CNTK Wiki - Github." Web page, 13-Feb-2017 [Online]. Available: <https://github.com/Microsoft/CNTK/wiki>

[186] Modine Austin, "Cobbler." Web page, Feb-2008 [Online]. Available:

http://www.theregister.co.uk/2008/06/19/red_hat_summit_2008_cobb

[187] R. Cilibarsi, A. L. Cruz, S. de Rooij, and M. Keijzer, "What is complearn." Web page [Online]. Available: <http://complearn.org/>

[188] CoreOS, "Why coreos." Web page, Jan-2017 [Online]. Available: <https://coreos.com/why/>

[189] R. Grehan, "NoSQL showdown: MongoDB vs. Couchbase." Web page, Mar-2013 [Online]. Available: <http://www.infoworld.com/article/2613970/nosql/nosql-showdown--mongodb-vs--couchbase.html>

[190] M. Brown, "The technology behind couchbase." Web page, Mar-2012 [Online]. Available: <https://www.safaribooksonline.com/blog/2012/03/01/the-technology-behind-couchbase/>

[191] Wikipedia, "Erlang (programming language)." Web page, Jan-2017 [Online]. Available: [https://en.wikipedia.org/wiki/Erlang_\(programming_language\)](https://en.wikipedia.org/wiki/Erlang_(programming_language))

[192] Erlang Central, "Couchbase performance and scalability: Iterating with dtrace observability." Web page, Mar-2012 [Online]. Available: <http://erlangcentral.org/videos/couchbase-performance-and-scalability-iterating-with-dtrace-observability/#.WI5uYephnRY>

[193] Sean Lynch, "Why Membase Uses Erlang." Web page, Oct-2010 [Online]. Available: [{https://blog.couchbase.com/why-membase-uses-erlang}](https://blog.couchbase.com/why-membase-uses-erlang)

[194] R. Kalla, "Well put! When should you use mongodb vs

couchbase versus redis..." Web page, Oct-2011 [Online]. Available: <http://rick-hightower.blogspot.com/2014/04/well-put-when-should-you-use-mongodb-vs.html>

[195] R. Smith, "What are the advantages and disadvantages of using mongodb vs couchdb vs cassandra vs redis?" Web page, Nov-2015 [Online]. Available: <https://www.quora.com/What-are-the-advantages-and-disadvantages-of-using-MongoDB-vs-CouchDB-vs-Cassandra-vs-Redis>

[196] J. Lennon, "Exploring couchdb: A document-oriented database for web applications." Web page, Mar-2009 [Online]. Available: <http://www.ibm.com/developerworksopensource/library/os-couchdb/index.html>

[197] Apache Software Foundation, "1.1. Document storage." Web page, Feb-2017 [Online]. Available: <http://docs.couchdb.org/en/stable/intro/overview.html>

[198] Couchbase, Inc., "Couchbase and apache couchdb compared." Web page, Feb-2017 [Online]. Available: <https://www.couchbase.com/couchbase-vs-couchdb>

[199]:o: FIX ENTRY, "CouchDB." Web page [Online]. Available: <http://docs.couchdb.org/en/master/intro/index.html>

[200] "What is json?" Web page [Online]. Available: <https://developers.squarespace.com/what-is-json>

[201] "CouchDB." Web page [Online]. Available: <https://media.readthedocs.org/pdf/couchdb/latest/couchdb.pdf>

[202] J. Lehnardt, "Couchdb the definitive introduction." Web page, Aug-2014 [Online]. Available: <https://www.infoq.com/articles/apache-couchdb-the-definitive-introduction>

[203] Arvados, "Arvados." Web Page [Online]. Available: <https://www.arvados.org/>

- [204] Arvados, "Arvados contributor wiki." Web Page [Online]. Available: <https://dev.arvados.org/>
- [205] "CUBRID." Web page, 2017 [Online]. Available: <http://www.cubrid.org/>
- [206] Singh, Dilpreet, Reddy, and C. K., "A survey on platforms for big data analytics," Journal of Big Data, vol. 2, no. 1, p. 8, Oct. 2014 [Online]. Available: <https://doi.org/10.1186/s40537-014-0008-6>
- [207] C. Lin, "E6895 big data analytics lecture 7." Presentation, 2016 [Online]. Available: <https://www.ee.columbia.edu/~cylin/course/bigdata/EECS6895-AdvancedBigDataAnalytics-Lecture7.pdf>
- [208] "CUDA programming." Web page, Jan-2017 [Online]. Available: <http://www.big-data.tips/cuda-programming>
- [209] Wikipedia, "CUDA." Web page, Feb-2017 [Online]. Available: <https://en.wikipedia.org/wiki/CUDA>
- [210] "D3 data-driven documents." Web page [Online]. Available: <https://d3js.org/>
- [211] Wikipedia, "Data analytics acceleration library." 2017 [Online]. Available: https://en.wikipedia.org/wiki/Data_Analytics_Acceleration_Library
- [212] "Intel data analytics acceleration library (intel daal)." Web page [Online]. Available: <https://software.intel.com/en-us/intel-daal>
- [213] I. D. Zone, "Intel data analytics acceleration library (intel daal)." Web page, 2017 [Online]. Available: <https://software.intel.com/en-us/intel-daal>
- [214] "DataFu." Web page [Online]. Available: <https://datafu.incubator.apache.org/>
- [215] "Apache datafu." Web page [Online]. Available:

<https://datafu.apache.org/>

[216] J. Hagadorn, S. O'Meara, J. Timberman, B. Berry, and N. Harvey, "Roles, environments, attributes, and data bags - part 2." Web page [Online]. Available: <http://foodfightshow.org/2013/01/roles2.html>

[217] DataNucleus, "DataNucleus support." Web page [Online]. Available: <http://www.datanucleus.com/>

[218] Wikipedia, "DataNucleus wikipedia." Web page, 2008 [Online]. Available: <https://en.wikipedia.org/wiki/DataNucleus>

[219] "Open source data turbine initiative." Web page [Online]. Available: <http://dataturbine.org/>

[220] "Programming for dataturbine." Webpage, 2017 [Online]. Available: <http://dataturbine.org/documentation/programming-for-dataturbine/>

[221] P. S. Tony Fountain Sameer Tilak, "The open source dataturbine initiative: Empowering the scientific community with streaming data middleware," The Bulletin of the Ecological Society of America, vol. 93, 2012 2012 [Online]. Available: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20090019753.pdf>

[222] "DB2 introduction." Web page, Jun-2016 [Online]. Available: https://www.tutorialspoint.com/db2/db2_introduction.htm

[223] Wikipedia, "IBM db2-wikipedia." Web page, Feb-2017 [Online]. Available: https://en.wikipedia.org/wiki/IBM_DB2

[224] DC.js, "Dc.js - dimensional charting javascript library." Web page, Jan-2017 [Online]. Available: <https://dc-js.github.io/dc.js/>

[225] Tutorials point, "DC.js tutorial." Web page [Online]. Available: <https://www.tutorialspoint.com/dcjs/>

[226] J. Wettinger, U. Breitenbücher, and F. Leymann, "DevOpSlang - bridging the gap between development and operations," in

Proceedings of the 3rd european conference on service-oriented and cloud computing (esocc 2014), 2014, pp. 108–122.

[227] S. Pooya, “Discoproject.” Web page, Dec-2016 [Online]. Available: <https://github.com/discoproject/disco>

[228] The Disco Project, “What is disco.” Web page, Feb-2017 [Online]. Available: <http://disco.readthedocs.io/en/develop/intro.html>

[229] The Disco Project, “Why not hadoop?” Web page, Feb-2017 [Online]. Available: <http://disco.readthedocs.io/en/develop/faq.html#why-not-hadoop>

[230] T. Clarridge, “Disco - a powerful erlang and python map/reduce framework.” Blog, May-2014 [Online]. Available: <http://www.taitclarridge.com/techlog/2014/05/disco-a-powerful-erlang-and-python-mapreduce-framework.html>

[231] Nokia Corp., “Nokia.” Web page, Feb-2017 [Online]. Available: http://www.nokia.com/en_int

[232] Wikipedia, “Deeplearning4j.” Web page, Feb-2017 [Online]. Available: <https://en.wikipedia.org/wiki/Deeplearning4j>

[233] A. Kumar, “Docker reference architecture: Universal control plane 3.0 service discovery and load balancing.” Web page [Online]. Available: <https://success.docker.com/article/ucp-service-discovery>

[234] B. Bashari Rad, H. Bhatti, and M. Ahmadi, “An introduction to docker and analysis of its performance,” IJCSNS International Journal of Computer Science and Network Security, 2017 [Online]. Available: <https://www.researchgate.net/publication/318816158/download>

[235] M. List, “Using docker compose for the simple deployment of an integrated drug target screening platform,” Journal of Integrative Bioinformatics, vol. 14, no. 2, Jun. 2017 [Online]. Available: <https://doi.org/10.1515/jib-2017-0016>

[236] K. Matthias and S. P. Kane, Docker up and running, shipping

reliable containers in production. 1005 Gravenstein Highway North, Sebastopol, CA 95472.: O'REILLY, 2015.

[237] "Docker swarm." Web page [Online]. Available: <https://www.docker.com/products/docker-swarm>

[238] Dokku, "WELCOME to dokku!" Git Hub, Apr-2016 [Online]. Available: <https://dokku.github.io/first-prost/welcome-to-dokku>

[239] Dokku, "View documentation." Web Page, Apr-2016 [Online]. Available: <http://dokku.viewdocs.io/dokku/>

[240] Dokku, "Getting started with dokku." Web Page, Apr-2016 [Online]. Available: <http://dokku.viewdocs.io/dokku/getting-started/installation/>

[241] J. Novet, "DotCloud." Web page, Jan-2016 [Online]. Available: <http://venturebeat.com/2016/01/22/dotcloud-the-cloud-service-that-gave-birth-to-docker-is-shutting-down-on-february-29/>

[242] Vodafone, "Help fight cancer tonight with the dreamlab app." Web page, 2017 [Online]. Available: <https://www.vodafone.com.au/foundation/dreamlab>

[243] A. Jones, "5 minute read: HOW does dreamlab work." Web page, 2017 [Online]. Available: <https://www.vodafone.com.au/red-wire/dreamlab>

[244] J. Barr, "Vodafone dreamLab accelerating cancer research." Web page, Feb-2016 [Online]. Available: <https://aws.amazon.com/blogs/aws/vodafone-dreamlab-accelerating-cancer-research/>

[245] T. A. S. Foundation, "Drill introduction." Aug-2018 [Online]. Available: <https://drill.apache.org/docs/drill-introduction/>

[246] T. A. S. Fondation, "Architecture introduction." [Online]. Available: <https://drill.apache.org/docs/architecture-introduction/>

- [247] T. A. S. Fondation, "Drill query execution." [Online]. Available: <https://drill.apache.org/docs/drill-query-execution/>
- [248] Wikipedia, "Dryad(Programming)-wikipedia." Web page, Nov-2016 [Online]. Available: [https://en.wikipedia.org/wiki/Dryad_\(programming\)](https://en.wikipedia.org/wiki/Dryad_(programming))
- [249] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks," in ACM sigops operating systems review, 2007, vol. 41, pp. 59-72 [Online]. Available: <https://www.microsoft.com/en-us/research/wp-content/uploads/2007/03/eurosys07.pdf>
- [250] Microsoft, "Dryad." Web page, Dec-2014 [Online]. Available: <https://www.microsoft.com/en-us/research/project/dryad/>
- [251] H. Hiden, S. Woodman, P. Watson, and J. Cala, "Developing cloud applications using the e-science central platform," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 371, no. 1983, Dec. 2012.
- [252] Wikipedia, "EclipseLink." 2010 [Online]. Available: <https://en.wikipedia.org/wiki/EclipseLink>
- [253] L. Florio and K. Wierenga, "Eduroam, providing mobility for roaming users," TERENA, 2005 [Online]. Available: <https://www.terena.org/activities/tf-mobility/docs/ppt/eunis-eduroamfinal-LF.pdf>
- [254] "EDUROAM." Web page [Online]. Available: <https://www.eduroam.org/about/>
- [255] "Ehcache - features." Web page [Online]. Available: <http://www.ehcache.org/about/features.html>
- [256] "Ehcache - documentation." Web page [Online]. Available: <http://www.ehcache.org/documentation/3.2/getting-started.html>
- [257] S. Banon, "Elastic search." 2018 [Online]. Available:

<https://en.wikipedia.org/wiki/Elasticsearch>

[258] S. Banon, “Elastic search.” 2018 [Online]. Available: <https://aws.amazon.com/elasticsearch-service/what-is-elasticsearch/>

[259] S. Banon, “Elastic search.” 2018 [Online]. Available: <https://github.com/elastic/elasticsearch>

[260] S. Banon, “Elastic search.” 2018 [Online]. Available: <https://qbox.io/blog/what-is-elasticsearch>

[261] D. Sullivan, “Paas provider comparison guide: Engine yard - orchestration and management,” Cloud Computing. Article; tomsitpro.com, Jul-2013 [Online]. Available: <http://www.tomsitpro.com/articles/paas-engine-yard-amazon-elastic-cloud-computing,2-578.html>

[262] A. Auradkar, “Introducing espresso - linkedin’s hot new distributed document store.” Website, Jan-2015 [Online]. Available: <https://engineering.linkedin.com/espresso/introducing-espresso-linkedin-s-hot-new-distributed-document-store>

[263] L. Qiao et al., “On brewing fresh espresso: LinkedIn’s distributed data serving platform.” in SIGMOD conference, 2013, pp. 1135–1146 [Online]. Available: <http://dblp.uni-trier.de/db/conf/sigmod/sigmod2013.html#QiaoSDQSGCSZABBGGIJLF>

[264] D. Nurmi et al., “The eucalyptus open-source cloud-computing system,” in Proceedings of the 2009 9th ieee/acm international symposium on cluster computing and the grid, 2009, pp. 124–131.

[265] “The eucalyptus open-source private cloud.” Web page [Online]. Available: <http://www.cloudbook.net/resources/stories/the-eucalyptus-open-source-private-cloud>

[266] Microsoft, “Event hubs.” Web Page [Online]. Available: <https://azure.microsoft.com/en-us/services/event-hubs/>

[267] S. Vijayasarathy, S. Pelluru, S. Ahuja, and S. Manheim, “What is

azure event hubs?" Web Page [Online]. Available: <https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-about>

[268] Microsoft, "Azure event hubs documentation." Web Page [Online]. Available: <https://docs.microsoft.com/en-us/azure/event-hubs/>

[269] S. Muralidhar et al., "F4: Facebook's warm blob storage system," in Proceedings of the 11th usenix conference on operating systems design and implementation, 2014, pp. 383-398.

[270] Facebook Inc., "Under the hood: Scheduling mapreduce jobs more efficiently with corona." Web page, Nov-2012 [Online]. Available: <https://www.facebook.com/notes/facebook-engineering/under-the-hood-scheduling-mapreduce-jobs-more-efficiently-with-corona/10151142560538920/>

[271] "Facebook's new realtime analytics system: HBase to process 20 billion events per day." Web page, Mar-2011 [Online]. Available: <http://highscalability.com/blog/2011/3/22/facebook-s-new-realtime-analytics-system-hbase-to-process-20.html>

[272] G. J. Chen et al., "Realtime data processing at facebook," in Proceedings of the 2016 international conference on management of data, 2016, pp. 1087-1098 [Online]. Available: <http://doi.acm.org/10.1145/2882903.2904441>

[273] Jordan Novet, "Facebook matured." Web page, Jun-2013 [Online]. Available: <https://gigaom.com/2013/06/25/how-facebook-matured-its-data-structure-and-stepped-into-the-graph-world/>

[274] Facebook, Inc., "TAO: The power of the graph." Web page, Jul-2013 [Online]. Available: <https://www.facebook.com/notes/facebook-engineering/tao-the-power-of-the-graph/10151525983993920/>

[275] Prashanth, "Facebook tupperware." Blog, Dec-2015 [Online]. Available: <http://blog.cspp.in/index.php/2015/12/17/facebook-tupperware/>

[tupperware/](#)

[276] Wikipedia, “Fiddler software.” Web page, Feb-2017 [Online]. Available: [https://en.wikipedia.org/wiki/Fiddler_\(software\)](https://en.wikipedia.org/wiki/Fiddler_(software))

[277] “FITS nasa.” Web page [Online]. Available: <https://fits.gsfc.nasa.gov/>

[278] “FITS news.” Web page [Online]. Available: https://fits.gsfc.nasa.gov/fits_standard.html

[279] “FITS vatican library.” Web page [Online]. Available: <https://www.vatlib.it/home.php?pag=digitalizzazione&ling=eng>

[280] “Fits matlab.” Web page [Online]. Available: https://www.mathworks.com/help/matlab/import_export/importing-flexible-image-transport-system-fits-files.html?requestedDomain=www.mathworks.com

[281] W. K et al., Astronomical image processing with hadoop, vol. 442. Astronomical Data Analysis Software; Systems XX. ASP Conference Proceedings, 2011 [Online]. Available: <http://adsabs.harvard.edu/abs/2011ASPC..442...93W>

[282] ApacheFlink, “What is apache flink?” Web page, Oct-2018 [Online]. Available: <https://flink.apache.org/flink-applications.html>

[283] The Apache Software Foundation, “Apache flume.” Web page [Online]. Available: <https://flume.apache.org/index.html>

[284] IBM, “What is flume?” Web page [Online]. Available: <https://www-01.ibm.com/software/data/infosphere/hadoop/flume/>

[285] “FTP wikipedia.” Web page [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/RCFileCat>

[286] A. Bhushon, “A file transfer protocol,” Apr. 1971 [Online]. Available: <https://tools.ietf.org/html/rfc114>

[287] “FUSE site.” Web page [Online]. Available: <http://fuse.sourceforge.net>

[288] X. Zhang, D. Du, J. Hughes, R. Kavuri, and S. StorageTek, “Hptfs: A high performance tape file system,” in Proceedings of 14th nasa goddard/23rd ieee conference on mass storage system and technologies, 2006 [Online]. Available: https://www.dtc.umn.edu/publications/reports/2006_11.pdf

[289] T. Xu, K. Sato, and S. Matsuoka, “CloudBB: Scalable i/o accelerator for shared cloud storage,” in 2016 ieee 22nd international conference on parallel and distributed systems (icpads), 2016, pp. 509–518.

[290] “About ansible galaxy.” Web page [Online]. Available: <https://galaxy.ansible.com/intro/>

[291] M. Heap, Ansible from beginner to pro. apress, 2016.

[292] “Github ansible/ galaxy.” Web page [Online]. Available: <https://github.com/ansible/galaxy/>

[293] “Galera cluster.” Web page, 2017 [Online]. Available: <http://galeracluster.com/>

[294] I. team -University of Texas, “Galois website.” Web page [Online]. Available: <http://iss.ices.utexas.edu/?p=projects/galois>

[295] K. Pingali et al., “The tao of parallelism in algorithms,” in The tao of parallelism in algorithms, 2011, pp. 1–14 [Online]. Available: <http://iss.ices.utexas.edu/Publications/Papers/pingali11.pdf>

[296] “Ganglia monitoring system.” Web page [Online]. Available: <http://ganglia.info/>

[297] “Ganglia (software).” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Ganglia_\(software\)](https://en.wikipedia.org/wiki/Ganglia_(software))

[298] IU, “What is gffs.” Web page, Jan-2018 [Online]. Available:

<https://kb.iu.edu/d/bblz>

[299] Genesis II Wiki, “GFFS.” Web page, 2018 [Online]. Available: <http://genesis2.virginia.edu/wiki/Main/GFFS>

[300] X. Shi et al., “GIRAFFE: A scalable distributed coordination service for large-scale systems,” in GIRAFFE: A scalable distributed coordination service for large-scale systems, 2014, pp. 1–10 [Online]. Available: <http://www.mcs.anl.gov/papers/P5157-0714.pdf>

[301] Wikipedia, “Graph theory.” Web page, Oct-2018 [Online]. Available: https://en.wikipedia.org/wiki/Graph_theory

[302] M. Campf, “How-to: Write and run apache giraph jobs on apache hadoop.” Web page, Feb-2014 [Online]. Available: <http://blog.cloudera.com/blog/2014/02/how-to-write-and-run-giraph-jobs-on-hadoop/>

[303] S. Sakr, “Processing large-scale graph data: A guide to current technology.” Web page, Jun-2013 [Online]. Available: <https://www.ibm.com/developerworks/library/os-giraph/index.html>

[304] R. Goyal, “Introduction to apache giraph.” Video, Dec-2015 [Online]. Available: <https://www.youtube.com/watch?v=FzR5zoEN9n8>

[305] J. Lindsay, “Gitreceive.” Web page, Feb-2016 [Online]. Available: <https://github.com/programm/gitreceive>

[306] toolkit, “GT 6.0 gridftp.” web page [Online]. Available: <http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/>

[307] B. Sotomayor and L. Childers, Globus toolkit 4: Programming java services. Morgan Kaufmann, 2006.

[308] I. Foster, “Globus toolkit version 4: Software for service-oriented systems,” Journal of computer science and technology, vol. 21, no. 4, p. 513, 2006.

[309] “About the globus toolkit.” Web page [Online]. Available:

<http://toolkit.globus.org/toolkit/about.html>

[310] Alphabet, Inc., "Google cloud." Web page, Jan-2017 [Online]. Available: <https://cloud.google.com>

[311] "Google cloud platform." Web page [Online]. Available: <https://cloud.google.com>

[312] "What is a public cloud?" Web page [Online]. Available: <http://www.interoute.com/cloud-article/what-public-cloud>

[313] "AppEngine - platform as a service." Web page [Online]. Available: <https://cloud.google.com/appengine>

[314] Wikipedia, "Google app engine." Web page, Jan-2017 [Online]. Available: https://en.wikipedia.org/wiki/Google_App_Engine

[315] "What is bigquery? BigQuery documentation google cloud platform." Web page [Online]. Available: <https://cloud.google.com/bigquery/what-is-bigquery>

[316] "What is bigquery? Google cloud platform." Web page; google.com [Online]. Available: <https://cloud.google.com/bigquery>

[317] M. Pasumansky, "Inside capacitor, bigquery's next-generation columnar storage format." Blog, Apr-2016 [Online]. Available: <https://cloud.google.com/blog/big-data/2016/04/inside-capacitor-bigquerys-next-generation-columnar-storage-format>

[318] L. Thomson, "Google opens bigquery for cloud analytics." Web page, Nov-2011 [Online]. Available: https://www.theregister.co.uk/Print/2011/11/14/google_bigquery_clou

[319] Google, "BigQuery client libraries." Web page, Sep-2018 [Online]. Available: <https://cloud.google.com/bigquery/docs/reference/libraries>

[320] K. Sato, "An inside look at google bigquery," Google Whitepapers, Sep. 2012 [Online]. Available: <https://cloud.google.com/files/BigQueryTechnicalWP.pdf>

- [321] I. Google, “Cloud bigtable.” Online [Online]. Available: <https://cloud.google.com/bigtable/>
- [322] M. Burrows et al., “USENIX symposium on operating systems design and implementation.” Google, Inc., 2006 [Online]. Available: <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/68a74a85e1662fe02ff3967497f31fda7f32225c.pdf>
- [323] P. Krzyzanowski, “BigTable: A nosql massively parallel table.” Online, Nov-2011 [Online]. Available: <https://www.cs.rutgers.edu/~pxk/417/notes/content/bigtable.html>
- [324] A. Ailijiang, A. Charapko, and M. Demirbas, “Consensus in the cloud: Paxos systems demystified,” in 2016 25th international conference on computer communication and networks (icccn), 2016, pp. 1–10 [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/7568499/>
- [325] Google, “CLOUD dataflow.” Web page [Online]. Available: <https://cloud.google.com/dataflow/>
- [326] J. Jackson, “Google service analyzes live streaming data.” Web page, Jun-2014 [Online]. Available: <http://www.infoworld.com/article/2607938/data-mining/google-service-analyzes-live-streaming-data.html>
- [327] Google, “Cloud dataflow.” Web page [Online]. Available: <https://cloud.google.com/dataflow/>
- [328] E. McNulty, “What is google cloud dataflow?” Web page [Online]. Available: <http://dataconomy.com/2014/08/google-cloud-dataflow/>
- [329] Google Developers, “Cloud machine learning.” Web page [Online]. Available: <https://cloud.google.com/ml-engine/>
- [330] Google Developers, “Cloud ml engine overview.” Web page [Online]. Available: <https://cloud.google.com/ml-engine/docs/concepts/technical-overview>

[331] “Google cloud sql.” Web page [Online]. Available: <https://cloud.google.com/sql/>

[332] mysql, “MySQL 5.7 reference manual, what is mysql.” Web page [Online]. Available: <https://dev.mysql.com/doc/refman/5.7/en/what-is-mysql.html>

[333] “PostgreSQL wiki.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/PostgreSQL>

[334] “Google cloud sql faq.” Web page [Online]. Available: <https://cloud.google.com/sql/faq#version>

[335] “Google cloud storage documentation.” Web page [Online]. Available: <https://cloud.google.com/storage/docs>

[336] Google Inc., “Google cloud datastore overview.” Web page [Online]. Available: <https://cloud.google.com/datastore/docs/concepts/overview>

[337] Google Inc., “Balancing strong and eventual consistency with google cloud datastore.” Web page [Online]. Available: <https://cloud.google.com/datastore/docs/articles/balancing-strong-and-eventual-consistency-with-google-cloud-datastore/>

[338] S. Melnik et al., “Dremel: Interactive analysis of web-scale datasets,” Communications of the ACM, vol. 54, pp. 114–123, Jun. 2011 [Online]. Available: <http://cacm.acm.org/magazines/2011/6/108648-dremel-interactive-analysis-of-web-scale-datasets/fulltext>

[339] J. Shute et al., “F1: A distributed sql database that scales,” Proc. VLDB Endow., vol. 6, no. 11, pp. 1068–1079, Aug. 2013 [Online]. Available: <https://static.googleusercontent.com/media/research.google.com/en/>

[340] J. Shute et al., “F1: The fault-tolerant distributed rdbms supporting google’s ad business,” in Proceedings of the 2012 acm

sigmod international conference on management of data, 2012, pp. 777–778 [Online]. Available: <http://doi.acm.org/10.1145/2213836.2213954>

[341] C. Chambers et al., “FlumeJava: Easy, efficient data-parallel pipelines,” in ACM sigplan conference on programming language design and implementation (pldi), 2010, pp. 363–375 [Online]. Available: <http://dl.acm.org/citation.cfm?id=1806638>

[342] C. Metz, “Google unleashes more big-data genius with a new cloud service.” Web page, Jun-2014 [Online]. Available: <https://www.wired.com/2014/06/google-cloud-data-flow/>

[343] Nolan, “History of hadoop: MapReduce and flumejava.” Blog, Jul-2014 [Online]. Available: <https://www.bp-3.com/blog/history-of-distributed-computing-hadoop-mapreduce-and-flumejava/>

[344] Google, “Google fusion tables.” Web Page [Online]. Available: <https://developers.google.com/fusiontables/>

[345] Wikipedia, “Google fusion tables.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Google_Fusion_Tables

[346] A. Halevy, “Google fusion tables.” Web page, Jun-2009 [Online]. Available: <https://ai.googleblog.com/2009/06/google-fusion-tables.html>

[347] Google, “Google fusion tables: Data management, integration and collaboration in the cloud.” Google Inc., 2012 [Online]. Available: <http://homes.cs.washington.edu/~alon/files/socc10.pdf>

[348] kubernetes.io, “What is kubernetes?” Web page, Sep-2018 [Online]. Available: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>

[349] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, “Borg, omega, and kubernetes,” Queue, vol. 14, no. 1, pp. 10:70–10:93, Jan. 2016 [Online]. Available:

<http://doi.acm.org/10.1145/2898442.2898444>

[350] redhat.io, "What is kubernetes." Web page, Oct-2018 [Online]. Available: <https://www.redhat.com/en/topics/containers/what-is-kubernetes>

[351] grodrigues3, "Kubernetes design and architecture." Web page, Aug-2017 [Online]. Available: <https://github.com/kubernetes/community/blob/master/contributors/proposals/architecture/architecture.md>

[352] T. Akidau et al., "MillWheel: Fault-tolerant stream processing at internet scale," in Very large data bases, 2013, pp. 734–746.

[353] Google, "Google cloud prediction api documentation." Web page [Online]. Available: <https://cloud.google.com/prediction/docs/>

[354] Google, "Google cloud translation api documentation." Web page [Online]. Available: <https://cloud.google.com/translate/docs/>

[355] Google Inc., "What-is-google-pub-sub." Web page [Online]. Available: <https://cloud.google.com/pubsub/docs/overview>

[356] Google Inc., "Google-pub-sub-scalable-messaging-middleware." Web page [Online]. Available: <https://cloud.google.com/pubsub/>

[357] "Gora - in-memory data model and persistence for big data." Web page [Online]. Available: <http://gora.apache.org/>

[358] P. Shriddeep, B. Thilina, M. Matthew, and S. Ryan, "Granules." Project, Jul-2016 [Online]. Available: <http://granules.cs.colostate.edu/>

[359] "Intel graph analytics solutions: Intel graph builder for apache hadoop* software v2." Web page [Online]. Available: <https://www-ssl.intel.com/content/www/us/en/software/intel-graph-builder-for-apache-hadoop-software-v2-product-detail.html>

[360] A. Kyrola, G. Blelloch, and C. Guestrin, "GraphChi: Large-scale graph computation on just a pc," in Proceedings of the 10th usenix

conference on operating systems design and implementation, 2012, pp. 31–46 [Online]. Available: <http://dl.acm.org/citation.cfm?id=2387880.2387884>

[361] Wikipedia, “Graph database.” Web page, Feb-2017 [Online]. Available: https://en.wikipedia.org/wiki/Graph_database

[362] turi, “Turi (formerly graphlab).” Web page, Nov-2018 [Online]. Available: <https://turi.com/>

[363] J. Shieber, Web page, 2014 [Online]. Available: <https://techcrunch.com/2015/01/08/machine-learning-startup-graphlab-gets-a-new-name-and-an-18-5m-check/>

[364] B. Lorica, “The evolution of graphlab.” Web page, Jan-2015 [Online]. Available: <https://www.oreilly.com/ideas/the-evolution-of-graphlab>

[365] S. Ray, “Tutorial - getting started with graphlab for machine learning in python.” Web page, Dec-2015 [Online]. Available: <https://www.analyticsvidhya.com/blog/2015/12/started-graphlab-python/>

[366] Apache Software Foundation, “Apache spark graphx.” Web page, Feb-2017 [Online]. Available: <http://spark.apache.org/graphx/>

[367] A. Lurie, “Top 47 log management tools.” Web page, May-2014 [Online]. Available: <http://blog.profitbricks.com/top-47-log-management-tools>

[368] L. Galea, “Graylog2 optimization for high-log environments.” Web page, Jul-2012 [Online]. Available: <https://dzone.com/articles/graylog2-optimization-high-log>

[369] “Dashboards - 2.2.1 documentation.” Web page, 2012 [Online]. Available: <http://docs.graylog.org/en/2.2/pages/dashboards.html>

[370] “H-store.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/H-Store>

- [371] A. Pavlo, J. Arulraj, and L. Ma, “H-store.” Web page [Online]. Available: <https://db.cs.cmu.edu/projects/h-store/>
- [372] R. Kallman et al., “H-store: A high-performance, distributed main memory transaction processing system.” Paper [Online]. Available: <http://www.vldb.org/pvldb/1/1454211.pdf>
- [373] “H2O Website Documentation components.” Web page [Online]. Available: <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html>
- [374] D. Cook, Practical machine learning with h2o. O'Reilley Media, 2017, p. 300 [Online]. Available: <https://books.google.com/books?id=nJWmDQAAQBAJ&pg=PP2&dq=h2o+software>
- [375] “H2O Wikipedia Documentation components.” Web page [Online]. Available: [https://en.wikipedia.org/wiki/H2O_\(software\)](https://en.wikipedia.org/wiki/H2O_(software))
- [376] Wikipedia, “Apache hadoop.” Web page, Mar-2017 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop
- [377] E. Coppa, “Hadoop architecture overview.” Code Repository [Online]. Available: <http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview>
- [378] B. Pawlikowski, A. Abouzeid, D. Abadi, and A. Silberschatz, “An architectural hybrid of mapreduce and dbms technologies for analytical workloads.” Web page [Online]. Available: db.cs.yale.edu/hadoopdb/hadoopdb.html
- [379] Apache, “Apache hama.” Web page [Online]. Available: <https://hama.apache.org/>
- [380] J. J. (Jong Hyuk) Park, H. Jin, Y.-S. Jeong, and M. K. Khan, Advanced multimedia and ubiquitous engineering. Springer, 2016.
- [381] Harp, “Harp.” Web page, Jan-2012 [Online]. Available: <http://harpjs.com>

[382] “Haystack.” Web page [Online]. Available: www.project-haystack.org

[383] Wikipedia, “Hazelcast.” Web page, Jan-2017 [Online]. Available: <https://en.wikipedia.org/wiki/Hazelcast>

[384] Hazelcast, “Open source in-memory data grid.” Code Repository, Jan-2017 [Online]. Available: <https://github.com/hazelcast/hazelcast>

[385] “Apache hbase.” Web page [Online]. Available: <https://hbase.apache.org/>

[386] D. Flair, “HBase use cases and real time applications 2018.” Webpage, Jun-2018 [Online]. Available: <https://data-flair.training/blogs/hbase-use-cases/>

[387] “HBase architecture, data flow, and use cases.” Webpage [Online]. Available: <https://www.guru99.com/hbase-architecture-data-flow-usecases.html>

[388] A. Sharma, “Apache hbase: Overview and use cases.” Webpage - PDF [Online]. Available: <https://events.static.linuxfound.org/sites/events/files/slides/ApacheBigData.pdf>

[389] M. Bertozzi, “How scaling really works in apache hbase.” Webpage, Apr-2013 [Online]. Available: <http://blog.cloudera.com/blog/2013/04/how-scaling-really-works-in-apache-hbase/>

[390] T. A. S. Fondation, “HCatalog usinghcat.” Aug-2013 [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/HCatalog+UsingHCat>

[391] W. Rowe, “What is apache hcatalog? HCatalog explained.” Aug-2017 [Online]. Available: <https://www.bmc.com/blogs/what-is-apache-hcatalog-hcatalog-explained/>

[392] Hortonworks Inc., “Hortonworks dataflow (hdf).” Web page [Online]. Available: <https://hortonworks.com/products/dataflow/>

[platforms/hdf/](#)

[393] “Hdfs.” Web page [Online]. Available: <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>

[394] “Apache helix.” Web page [Online]. Available: <https://helix.apache.org/>

[395] T. Teofili, “Powered by apache uima.” Web page, Dec-2015 [Online]. Available: <https://cwiki.apache.org/confluence/display/UIMA/Powered+by+Apac>

[396] C. Wodehouse, “Launch, run, and scale your application in the cloud with heroku.” Web page [Online]. Available: <https://www.upwork.com/hiring/development/launch-run-and-scale-your-application-in-the-cloud-with-heroku/>

[397] R. Chauvel, “Common challenges with big data deployments.” Blog, Jun-2016 [Online]. Available: <https://www.bmc.com/blogs/common-challenges-with-big-data-deployments/>

[398] D. Nield, “Cloud tech is helping small firms tap into big data.” The telegraph, Jul-2017 [Online]. Available: <https://www.telegraph.co.uk/connect/small-business/tech/cloud-tech-helping-small-firms-tap-big-data/>

[399] “Big data on heroku hadoop from treasure data.” Blog, Aug-2012 [Online]. Available: <https://bighadoop.wordpress.com/2012/08/23/big-data-on-heroku-hadoop-from-treasure-data/>

[400] I. Szegedi, “Big data on heroku-treasure data hadoop.” Web page, Aug-2012 [Online]. Available: <https://dzone.com/articles/big-data-heroku-treasure-data>

[401] V. B, S. YV, S. G, and Priya3, “HIBERNATE technology for an

efficient business application extension:" Paper, Jun-2011 [Online]. Available: <https://www.rroij.com/open-access/hibernate-technology-for-an-efficient-business-application-extension-118-125.pdf>

[402] Apache Hive, "APACHE hive tm." Web page, Nov-2018 [Online]. Available: <https://hive.apache.org/>

[403] intelliPaat, "What is apache hive?" Web page, Jun-2018 [Online]. Available: <https://intellipaat.com/blog/what-is-apache-hive/>

[404] M. Rouse, "Apache hive." Web page, Jul-2018 [Online]. Available: <https://searchdatamanagement.techtarget.com/definition/Apache-Hive>

[405] A. Verma, "Apache hive - a faster and better sql on hadoop." Web page, Jan-2018 [Online]. Available: <https://www.whizlabs.com/blog/apache-hive-faster-better-sql-on-hadoop/>

[406] "High performance parallel (hpx-5)." Web page [Online]. Available: <https://hpx.crest.iu.edu/>

[407] HPX-5: User guide. HPX-5; Web page [Online]. Available: https://hpx.crest.iu.edu/users_guide

[408] "What is htcondor?" Web page [Online]. Available: <https://research.cs.wisc.edu/htcondor/description.html>

[409] "What is hubzero?" Web page [Online]. Available: <https://hubzero.org/documentation/1.0.0/installation>

[410] M. McLennan and R. Kennell, "HUBzero: A platform for dissemination and collaboration in computational science and engineering," Computing in Science Engineering, vol. 12, no. 2, pp. 48-53, Mar. 2010.

[411] Wikipedia, "Hypervisor." Web page, Feb-2017 [Online]. Available: <https://en.wikipedia.org/wiki/Hypervisor>

- [412] Microsoft, "Introduction to hyper-v on windows 10." Website, Jun-2018 [Online]. Available: <https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/about/>
- [413] R. Corradini, "What is hyper-v?: The authoritative guide." Jun-2018 [Online]. Available: <https://www.5nine.com/hyper-v-authoritative-guide/>
- [414] H. J. Sarah Cooley Justin, "Hyper-v architecture." Oct-2018 [Online]. Available: <https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/reference/hyper-v-architecture>
- [415] IBM, "IBM cloud (formerly ibm bluemix)." Online, Nov-2018 [Online]. Available: <https://console.bluemix.net/docs/overview/ibm-cloud.html#overview>
- [416] M. Rouse, "IBM cloud (formerly ibm bluemix and ibm softlayer)." Online, May-2017 [Online]. Available: <https://searchcloudcomputing.techtarget.com/definition/IBM-Bluemix>
- [417] B. Butler, "PaaS primer: What is platform as a service and why does it matter?" NetworkWorld, Feb. 2013 [Online]. Available: <https://www.networkworld.com/article/2163430/cloud-computing/paas-primer--what-is-platform-as-a-service-and-why-does-it-matter-.html>
- [418] N. H. Raj, "Beginner's guide to deploying your blockchain in ibm bluemix," Hackernoon, May 2018 [Online]. Available: <https://hackernoon.com/beginners-guide-to-deploying-your-blockchain-in-ibm-bluemix-da11d09f3914>
- [419] L. Sun, "IBM and microsoft are upgrading walmart's digital supply chain," Motley Fool, Sep. 2018 [Online]. Available: <https://www.fool.com/investing/2018/09/30/ibm-and-microsoft-are-upgrading-walmarts-digital-s.aspx>
- [420] Wikipedia, "IBM cloudant." Web page, Feb-2017 [Online]. Available: <https://en.wikipedia.org/wiki/Cloudant>

- [421] “5 things to know about dashDB.” Web page [Online]. Available: <https://www.ibm.com/developerworks/community/blogs/5things/entry?lang=en>
- [422] “IBM dashDB.” Web page [Online]. Available: <https://www.ibm.com/analytics/us/en/technology/cloud-data-services/dashdb/>
- [423] Wikipedia, “IBM general parallel file system.” Web page, Jan-2017 [Online]. Available: https://en.wikipedia.org/wiki/IBM_General_Parallel_File_System
- [424] IBM, “IBM spectrum scale.” Web page [Online]. Available: <http://www-03.ibm.com/systems/storage/spectrum/scale/>
- [425] IBM Corporation, “IBM system g documentation.” IBM; Web page, 2014 [Online]. Available: <http://systemg.research.ibm.com/>
- [426] IBM Corporation, “IBM system g documentation-2.” Predictive Analytics Today; Web page, 2017 [Online]. Available: <http://www.predictiveanalyticstoday.com/ibm-system-g-native-store/>
- [427] C. Y. Lin, “Graph computing and linked big data,” in 8th IEEE/ICSC international conference on semantic computing, 2014, pp. 1-63 [Online]. Available: <http://ieeexplore.ieee.org/iel5/7205220/7205221/7205221.pdf>
- [428] “IBM watson.” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Watson_\(computer\)](https://en.wikipedia.org/wiki/Watson_(computer))
- [429] “IBM watson product page.” Web page [Online]. Available: <https://www.ibm.com/watson>
- [430] “Wikipedia watson.” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Watson_\(computer\)](https://en.wikipedia.org/wiki/Watson_(computer))
- [431] ImageJ, “ImageJ introduction.” Web page, Feb-2017 [Online]. Available: <https://imagej.nih.gov/ij/docs/intro.html>

[432] Cloudera Inc., "Cloudera impala overview." Web page [Online]. Available: https://www.cloudera.com/documentation/enterprise/5-3-x/topics/impala_intro.html

[433] For Dummies, Inc., "Cloudera imapala and hadoop." Web page; John Wiley & Sons [Online]. Available: <http://www.dummies.com/programming/big-data/hadoop/cloudera-impala-and-hadoop/>

[434] ACADGILD, "Beginner's guide for impala." Web page, Mar-2016 [Online]. Available: <https://acadgild.com/blog/beginners-guide-impala/>

[435] J. C. Lizhe Wang Wei Jie, Grid computing: Infrastructure, service, and applications. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487: Taylor & Francis, 2009.

[436] "Inca - periodic, automated, user-level cyberinfrastructure testing." Web page [Online]. Available: <http://inca.sdsc.edu/>

[437] InCommon, "What is the incommon federation?" Web page [Online]. Available: https://spaces.internet2.edu/download/attachments/2764/final_InCon

[438] "Introduction to infinispan: Distributed in-memory key/value data grid and cache." Web page [Online]. Available: <http://infinispan.org/about/>

[439] "Infinispan." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Infinispan>

[440] iRods, "IRods." Web page [Online]. Available: <https://irods.org/>

[441] iRods, "IRods." Web page [Online]. Available: <https://github.com/irods/irods>

[442] A. jclouds, "What is jclouds?" Web page, Oct-2018 [Online]. Available: <https://jclouds.apache.org/start/what-is-jclouds/>

[443] Rackspace, "Jclouds is an apache top level project." Web page, Oct-2018 [Online]. Available: <https://developer.rackspace.com/blog/jclouds-is-an-apache-tlp/>

[444] H. Jayathilaka, "An introduction to apache jclouds, featuring the compute service api and blobstore." Web page, Oct-2018 [Online]. Available: <https://www.slideshare.net/hiranya911/jclouds>

[445] A. jclouds, "Jclouds core concepts." Web page, Oct-2018 [Online]. Available: <https://jclouds.apache.org/start/concepts/>

[446] "Jelastic," Wikipedia. Web page, 13-Feb-2017 [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Jelastic&oldid=754931676>

[447] "Grow Hosting Business with Software Platform," Jelastic. Web page, 13-Feb-2017 [Online]. Available: <https://jelastic.com/cloud-business-for-hosting-providers/>

[448] The Apache Software Foundation, "What is jena." Web page, Oct-2018 [Online]. Available: https://jena.apache.org/about_jena/about.html

[449] WC3, "OWL web ontology language guide." Web page, Nov-2009 [Online]. Available: <https://www.w3.org/TR/2004/REC-owl-guide-20040210/#Introduction>

[450] Integration made easy - jitterbit. 1 Kaiser Plaza Suite 701 Oakland, CA 94612: Jitterbit,Inc. [Online]. Available: <http://www.jitterbit.com/Files/Product/Jitterbit-General-Datasheet.pdf>

[451] Jitterbit, "Data integration and etl | jitterbit | integrate data from any source." Jitterbit; Web page [Online]. Available: <https://www.jitterbit.com/etl-data-integration/>

[452] Technical overview - jitterbit. Jitterbit,Inc, 2011 [Online]. Available: <http://www.jitterbit.com/Files/Product/JitterbitTechnicalOverview.pdf>

[453] Wikipedia, "Java message service - wikipedia." Web page

- [Online]. Available:
https://en.wikipedia.org/wiki/Java_Message_Service
- [454] “The java ee 6 tutorial.” Web page [Online]. Available:
<http://docs.oracle.com/javaee/6/tutorial/doc/bnkeh.html>
- [455] Jujucharms, Web page [Online]. Available:
<https://jujucharms.com/>
- [456] K. Subramanian, “Ubuntu ensemble is now juju.” Web page, Sep-2011 [Online]. Available:
<https://www.cloudave.com/14950/ubuntu-ensemble-is-now-juju/>
- [457] Wikipedia, “Juju (software).” Web page [Online]. Available:
[https://en.wikipedia.org/wiki/Juju_\(software\)](https://en.wikipedia.org/wiki/Juju_(software))
- [458] Jujucharms, Web page [Online]. Available:
<https://jujucharms.com/how-it-works>
- [459] DiscoverSDK, Web page [Online]. Available:
<http://www.discoversdk.com/products/juju#/overview>
- [460] “Project Jupyter.” Web page, 27-Feb-2017 [Online]. Available:
<http://www.jupyter.org>
- [461] “Github - jupyter/jupyter: Jupyter metapackage for installation, docs and chat.” Web page, 27-Feb-2017 [Online]. Available:
<https://github.com/jupyter/jupyter>
- [462] “Github - jupyter/notebook: Jupyter Interactive Notebook.” Web page, 27-Feb-2017 [Online]. Available:
<https://github.com/jupyter/notebook>
- [463] “The Jupyter notebook — Jupyter Notebook 5.0.0.dev documentation.” Web page, 27-Feb-2017 [Online]. Available:
<https://jupyter-notebook.readthedocs.io/en/latest/index.html>
- [464] “IPython,” Wikipedia. Web page, 27-Feb-2017 [Online]. Available:
<https://en.wikipedia.org/w/index.php?title=IPython&oldid=767266837>

[465] “What is the Jupyter Notebook? — Jupyter Notebook 5.0.0.dev documentation.” Web page, 27-Feb-2017.

[466] “Kafka.” Web Page [Online]. Available: <https://kafka.apache.org/intro>

[467] “Dzone.” Web Page [Online]. Available: <https://dzone.com/articles/what-is-kafka>

[468] G. von Laszewski and M. Hategan, “Using karajan,” in Java cog kit karajan/gridant workflow guide, 2005 [Online]. Available: <https://pdfs.semanticscholar.org/48e1/31ca4e7a76ca98c43d46975cd7>

[469] Kepler Project, “The kepler project.” Web page, Feb-2017 [Online]. Available: <https://kepler-project.org>

[470] Agargenta, “About kestrel.” Web page [Online]. Available: <https://github.com/twitter-archive/kestrel>

[471] E. R. Sparks, S. Venkataraman, T. Kaftan, M. J. Franklin, and B. Recht, “KeystoneML: Optimizing pipelines for large-scale advanced analytics,” arXiv preprint arXiv:1610.09451, 2016.

[472] A. Woodie, “Spark Gets New Machine Learning Framework: KeystoneML.” 27-Feb-2017 [Online]. Available: <https://www.datanami.com/2015/05/26/spark-gets-new-machine-learning-framework-keystoneml/>

[473] “Kibana: Explore, Visualize, Discover Data Elastic.” Web page, 27-Feb-2017 [Online]. Available: <https://www.elastic.co/products/kibana>

[474] “Github - elastic/kibana: Kibana analytics and search dashboard for Elasticsearch.” Web page, 27-Feb-2017 [Online]. Available: <https://github.com/elastic/kibana>

[475] “How To Use Logstash and Kibana To Centralize Logs On Ubuntu 14.04,” DigitalOcean. Web page, 27-Feb-2017 [Online]. Available: <https://www.digitalocean.com/community/tutorials/how-to-use-logstash-and-kibana-to-centralize-and-visualize-logs-on-ubuntu->

14-04

- [476] “Elasticsearch,” Wikipedia. Web page, 27-Feb-2017 [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Elasticsearch&oldid=767434249>
- [477] “Kibana,” Wikipedia. Web page, 27-Feb-2017 [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Kibana&oldid=767434139>
- [478] “Introduction Kibana User Guide [5.2] Elastic.” Web page, 27-Feb-2017 [Online]. Available: <https://www.elastic.co/guide/en/kibana/current/introduction.html>
- [479] DevTopics, “Kite.” Web page [Online]. Available: <http://www.devtopics.com/kite-obscure-programming-language-of-the-month/>
- [480] VoltDB, “VoltDB.” Web page [Online]. Available: <https://www.wired.com/2016/04/kites-coding-assitant-spots-errors-finds-better-open-source/>
- [481] “KVM wikipedia documentation.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Kernel-based_Virtual_Machine
- [482] “KVM webpage documentation.” Web page [Online]. Available: http://www.linux-kvm.org/page/Main_Page
- [483] F. labs, “Nijo cabinet: A straightforward implementation of dbm.” Web page [Online]. Available: <http://fallabs.com/kyotocabinet/>
- [484] “Tokyo cabinet.” Web page [Online]. Available: <http://fallabs.com/tokyocabinet/>
- [485] “Kyoto cabinet.” Web page [Online]. Available: <http://fallabs.com/kyotocabinet/>
- [486] Amazon, “AWSLambda.” Web page [Online]. Available: <https://aws.amazon.com/lambda/faqs/>

- [487] Amazon, "AWSLambdaEvent." Web page [Online]. Available: <http://docs.aws.amazon.com/lambda/latest/dg/invoking-lambda-function.html#intro-core-components-event-sources>
- [488] TechTarget, "Lightweight directory access protocol." Web page, Feb-2017 [Online]. Available: <http://searchmobilecomputing.techtarget.com/definition/LDAP>
- [489] LevelDB, "LevelDB." Web page, Feb-2017 [Online]. Available: leveldb.org
- [490] A. Libcloud, "About libcloud." Web page [Online]. Available: <https://libcloud.apache.org/about.html>
- [491] A. libcloud, "Python library for interacting with many of the popular cloud service providers using a unified api." Web page [Online]. Available: <https://libcloud.apache.org>
- [492] A. libcloud, "Welcome to apache libcloud's documentation!" Web page [Online]. Available: <https://libcloud.readthedocs.io/en/latest>
- [493] "Libvirt virtualization api." Web page [Online]. Available: <https://libvirt.org/>
- [494] T. Jones, "Anatomy of the libvirt virtualization." Web page, Jan-2010 [Online]. Available: <https://www.ibm.com/developerworks/library/l-libvirt/>
- [495] J. Shun and G. E. Blelloch, "Ligra: A lightweight graph processing framework for shared memory," in ACM sigplan notices - ppopp '13, 2013 [Online]. Available: <http://dl.acm.org/citation.cfm?id=2442530>
- [496] G. E. B. Julian Shun and L. Dhulipala, "Smaller and faster: Parallel processing of compressed graphs with ligra+," in ACM sigplan notices - ppopp '13DCC '15 proceedings of the 2015 data compression conference, 2015, pp. 403-412 [Online]. Available: <http://dl.acm.org/citation.cfm?id=2860198>
- [497] Wikipedia, "LinkedIn technical overview." Web page [Online].

Available: <https://cloud.google.com/ml-engine/docs/concepts/technical-overview>

[498] DeZyre, "How linkedin uses hadoop to leverage big data analytics." Web page [Online]. Available: <https://www.dezyre.com/article/how-linkedin-uses-hadoop-to-leverage-big-data-analytics/229>

[499] Quora, "LinkedIn database architecture." Web page [Online]. Available: <https://www.quora.com/What-is-LinkedIn-s-database-architecture-like>

[500] LinkedIn Developers, "Getting started with the rest api." Web page [Online]. Available: <https://developer.linkedin.com/docs/rest-api>

[501] K. Kolyshkin, "Virtualization in linux," 2006 [Online]. Available: <http://mirror.ihc.ru/download.openvz.org/doc/openvz-intro.pdf>

[502] H. Lee, "Virtualization basics: Understanding techniques and fundamentals," Indiana University [Online]. Available: <http://dsc.soic.indiana.edu/publications/virtualization.pdf>

[503] E. Reshetova, J. Karhunen, T. Nyman, and N. Asokan, "Security of os-level virtualization technologies," 2014 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-11599-3_5

[504] H. Potzl and M. E. Fiuczynski, "Linux-vserver: Resource efficient os-level virtualization," 2007 [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.798&rep=rep1&type=pdf#page=151>

[505] I. Gunaratne, "Evolution of the linux container." Web page, Sep-2016 [Online]. Available: <https://medium.com/containermind/evolution-of-linux-containers-and-future-6f2adc1d5086>

[506] L. Kotthoff, "LLAMA: Leveraging learning to automatically manage algorithms." 30-Apr-2014 [Online]. Available:

<https://arxiv.org/abs/1306.1031>

[507] H. Chu, “DEVOXX france.” 2013 [Online]. Available: <https://www.youtube.com/watch?v=Rx1-in-a1Xc>

[508] H. Chu, “Lightning memory-mapped database manager (lmdb).” Online, Dec-2015 [Online]. Available: <http://www.lmdb.tech/doc/>

[509] Elasticsearch, “Logstash introduction.” Web page, Feb-2017 [Online]. Available: <https://www.elastic.co/guide/en/logstash/current/introduction.html>

[510] M. R. Karim, “Searching and indexing with apache lucene.” Web page, Jan-2017 [Online]. Available: <https://dzone.com/articles/apache-lucene-a-high-performance-and-full-featured>

[511] LuceneTutorials.com, “Lucene tutorial.” Web page, 2018 [Online]. Available: <http://www.lucenetutorial.com/basic-concepts.html>

[512] C. Nutt, “Gamasutra - Amazon launches new, free, high-quality game engine: Lumberyard.” 27-Feb-2017 [Online]. Available: http://www.gamasutra.com/view/news/265425/Amazon_launches_new

[513] Gamefromscratch, “Hands On With Amazon’s Lumberyard Game Engine - YouTube.” 27-Feb-2017 [Online]. Available: <https://www.youtube.com/watch?v=FUDITTbt4qE>

[514] M. Vis, “What Is Lumberyard? (Lumberyard Tutorials Series #1) - YouTube.” 27-Feb-2017 [Online]. Available: <https://www.youtube.com/watch?v=Fxwo3KqSsUI>

[515] S. T. LLC, “About the lustre file system.” [Online]. Available: <http://lustre.org/about/>

[516] OpenSFS, “Lustre file system, version 2.4 released.” Oct-2018 [Online]. Available: <http://opensfs.org/press-releases/lustre-file-system-version-2-4-released/>

- [517] Wikipedia, “Lustre (file system).” [Online]. Available: [https://en.wikipedia.org/wiki/Lustre_\(file_system\)](https://en.wikipedia.org/wiki/Lustre_(file_system))
- [518] I. H. M. S, “Introduction to the lustre file system.” Apr-2015 [Online]. Available: <https://insidehpc.com/2015/04/introduction-to-the-lustre-file-system/>
- [519] “Linux containers.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/LXC>
- [520] “Why use lxc (linux containers) ?” Web page [Online]. Available: <http://www.jpablo128.com/why-use-lxc-linux-containers/>
- [521] “Linux containers in use.” Web page [Online]. Available: <http://www.infoworld.com/article/3072929/linux/containers-101-linux-containers-and-docker-explained.html>
- [522] Linux, “What’s lxc?” Web page, 2016 [Online]. Available: <https://linuxcontainers.org/lxc/introduction/>
- [523] E. Biederman, “LXC.” Web page, 2008 [Online]. Available: <https://en.wikipedia.org/wiki/LXC>
- [524] “The lxd container hypervisor.” Web page [Online]. Available: <https://www.ubuntu.com/containers/lxd>
- [525] S. Gruber, “LXD2.0: Introduction to lxd.” Web page [Online]. Available: <https://blog.ubuntu.com/2016/03/14/lxd-2-0-introduction-to-lxd>
- [526] M. Rouse, “Linux container hypervisor.” Web page, Jan-2018 [Online]. Available: <https://searchitoperations.techtarget.com/definition/LXD-Linux-container-hypervisor>
- [527] “Apache mahout.” Web page [Online]. Available: <http://mahout.apache.org/>
- [528] “Marionette collective webpage documentation.” Web page,

2016 [Online]. Available: <https://docs.puppet.com/mcollective/>

[529] B. H. Jianlong Zhong, "Medusa: Simplified graph processing on gpus," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1-11, Apr. 2013 [Online]. Available: http://pdcc.ntu.edu.sg/xtra/paper/2013ago/Medusa_TPDS13.pdf

[530] A. L. Cambridge, "The medusa applications environment." Web page [Online]. Available: <http://www.cl.cam.ac.uk/research/dtg/attarchive/medusa.html>

[531] J. C. Corbett et al., "Spanner: Google's globally distributed database," ACM Transactions on Computer Systems (TOCS), vol. 31, no. 3, p. 8, 2013 [Online]. Available: http://dl.acm.org/ft_gateway.cfm?id=2491245&type=pdf

[532] M. L. Braun, "Magastore, Spanner - distributed databases." Web page, Mar-2013 [Online]. Available: <http://blog.mikiobraun.de/2013/03/more-google-papers-megastore-spanner-voted-commits.html>

[533] Memcached, "About memcached." Web page [Online]. Available: <https://memcached.org/about>

[534] M. Chand, "Introduction to apache mesos." Web page, Jan-2018 [Online]. Available: <https://dzone.com/articles/introduction-to-apache-mesos>

[535] A. S. Tutorials, "Apache mesos introduction, architecture and working." Web page, Apr-2017 [Online]. Available: <https://data-flair.training/blogs/apache-mesos-tutorial/>

[536] T. Kraska, A. Talwalkar, J. Duchi, R. Griffith, M. Franklin, and M. Jordan, "MLbase: A distributed machine-learning system." 2013 [Online]. Available: <https://amplab.cs.berkeley.edu/wp-content/uploads/2013/01/dmx1.pdf>

[537] T. Kraska, "MLbase: A distributed machine learning system."

Youtube, Oct-2013 [Online]. Available:
<https://www.youtube.com/watch?v=W-WPcINo8v0>

[538] “Apache spark’s mllib.” Web page [Online]. Available:
<http://spark.apache.org/mllib/>

[539] X. Meng et al., “MLlib: Machine learning in apache spark,” CoRR, vol. abs/1505.06807, 2015 [Online]. Available:
<http://arxiv.org/abs/1505.06807>

[540] D. Albanese, R. Visintainer, S. Merler, S. Riccadonna, G. Jurman, and C. Furlanello, “Mlpy: Machine learning python,” CoRR, vol. abs/1202.6548, 2012 [Online]. Available:
<http://arxiv.org/abs/1202.6548>

[541] “MLPY documentation.” Web page, 2012 [Online]. Available:
<http://mlpy.sourceforge.net/docs/3.5/>

[542] AdaptiveComputing, “Moab cluster suite.” Web page [Online]. Available: <https://www.adaptivecomputing.com/products/>

[543] “MQTT.” Web page [Online]. Available: <http://mqtt.org/>

[544] “MQTT floodnet.” Web page [Online]. Available:
<http://mqtt.org/projects/floodnet>

[545] “MR-mpi.” Web page, 2017 [Online]. Available:
<http://mapreduce.sandia.gov/doc>

[546] E. Yoon, “MRQL - a sql on hadoop miracle.” Web page [Online]. Available: <http://www.hadoopsphere.com/2013/04/mrql-sql-on-hadoop-miracle.html>

[547] Apache Software Foundation, “Apache mrql.” Web page, Apr-2016 [Online]. Available: <https://mrql.incubator.apache.org/>

[548] L. Fegaras, “LanguageDescription - mrql wiki.” Web page [Online]. Available: <https://wiki.apache.org/mrql/LanguageDescription>

- [549] M. Kofler, The definite guide to mysql 5, Third Edition. Berkley, CA, USA: Apress, 2005 [Online]. Available: <http://ebooks.bharathuniv.ac.in/gdlc1/gdlc1/Computer%20Science%20/>
- [550] C. Arsenault, "The pros and cons of 8 popular databases." Web page, Apr-2017 [Online]. Available: <https://www.keycdn.com/blog/popular-databases/#2-MySQL>
- [551] "Nagios components." Web page [Online]. Available: <https://www.nagios.org/projects/>
- [552] D. Josephsen, Nagios: Building enterprise-grade monitoring infrastructures for systems and networks, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2013.
- [553] C. Issariyapat, P. Pongpaibool, S. Mongkolluksame, and K. Meesublak, "Using nagios as a groundwork for developing a better network monitoring system," in 2012 proceedings of picmet '12: Technology management for emerging technologies, 2012, pp. 2771–2777.
- [554] D. G. Murray, F. McSherry, R. Isaacs, M. Isard, P. Barham, and M. Abadi, "Naiad: A timely dataflow system," in Proceedings of the twenty-fourth acm symposium on operating systems principles, 2013, pp. 439–455 [Online]. Available: <http://doi.acm.org/10.1145/2517349.2522738>
- [555] C. Thekkath, "Naiad - microsoft research." Microsoft; Web page, Oct-2011 [Online]. Available: <https://www.microsoft.com/en-us/research/project/naiad/>
- [556] Community Grids Lab IU, "The naradabrokering project @ iu community grids laboratory." Pervasive Technology Labs at Indiana University; Web page, 501 N. MORTON ST, SUITE 224 BLOOMINGTON IN 47404, Nov-2009 [Online]. Available: <http://www.naradabrokering.org/>
- [557] G. Fox and S. Pallickara, "Deploying the naradabrokering

substrate in aiding efficient web and grid service interactions," Proceedings of the IEEE, vol. 93, no. 3, pp. 564–577, Mar. 2005.

[558] Community Grids Lab IU, "Some of the salient features in naradabrokering." Pervasive Technology Labs at Indiana University; Web page, 501 N. MORTON ST, SUITE 224 BLOOMINGTON IN 47404, Nov-2009 [Online]. Available: <http://www.naradabrokering.org/>

[559] Wikipedia, "Neo4j." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Neo4j>

[560] S. A. Raza, "Neo4j." Presentation/ Slides [Online]. Available: <https://www.slideshare.net/aliraza995/neo4j-graph-storage-27104408>

[561] Neo4j, "Chapter 4. Clustering." Web page [Online]. Available: <https://neo4j.com/docs/operations-manual/current/clustering/>

[562] Neo4j, "Causal cluster." Web page [Online]. Available: <https://neo4j.com/docs/operations-manual/current/clustering/>

[563] Neo4j, "Highly available cluster." Web page [Online]. Available: <https://neo4j.com/docs/operations-manual/current/clustering/high-availability/architecture/>

[564] "Amazon neptune." Web page [Online]. Available: <https://aws.amazon.com/neptune/>

[565] U. Edward Hartnett and R. RK, "Experience with an enhanced netCDF data model and interface for scientific data access," in 24th conference on iips, 2008.

[566] "NetCDF: Introduction and Overview." Web page, 13-Feb-2017 [Online]. Available: https://www.unidata.ucar.edu/software/netcdf/docs_rc/

[567] "Netty site." Web page [Online]. Available: <http://netty.io/>

[568] N. Maurer and M. Wolfthal, Netty in action, 1st ed. Greenwich, CT, USA: Manning Publications, 2015 [Online]. Available:

http://www.ebook.de/de/product/21687528/norman_maurer.netty_in

[569] The Apache Software Foundation, “Apache nifi.” Web page, Oct-2018 [Online]. Available: <https://nifi.apache.org/>

[570] A. Dokaeva, “How to make etl simple and intuitive with nifi.” Web page, Mar-2018 [Online]. Available: <https://issart.com/blog/how-to-make-etl-simple-and-intuitive-with-nifi/>

[571] A. Bridgwater, “NSA ‘nifi’ big data automation project out in the open.” Web page, Jul-2015 [Online]. Available: <https://www.forbes.com/sites/adrianbridgwater/2015/07/21/nsa-nifi-big-data-automation-project-out-in-the-open/#68cdd7dc55d6>

[572] “Nimbus.” Web Page [Online]. Available: <http://www.nimbusproject.org/doc/nimbus/faq/#what-is-nimbus>

[573] O. Mashayekhi, H. Qu, C. Shah, and P. Levis, “Scalable, fast cloud computing with execution templates.” Paper [Online]. Available: <https://hci.stanford.edu/cstr/reports/2016-02.pdf>

[574] “Nimbus and cloud computing meet star production demands.” Web Page [Online]. Available: <https://www.sciencedaily.com/releases/2009/04/090406083906.htm>

[575] Ninefold, “Ninefold news.” Web page, Nov-2015 [Online]. Available: <http://ninefold.com/news/>

[576] Wasserman, “Network workbench tool.” Web page [Online]. Available: https://www.researchgate.net/publication/228906060_Network_Work

[577] OCCI, “Open cloud computing interface.” Web page, Oct-2018 [Online]. Available: <http://occi-wg.org/about/>

[578] T. Metsch, “Open cloud computing interface - platform.” Web page, Sep-2016 [Online]. Available: <https://www.ogf.org/documents/GFD.227.pdf>

[579] “Open database connectivity.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Open_Database_Connectivity

[580] “Java database connectivity.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Java_Database_Connectivity

[581] “Business process execution language.” Web page [Online]. Available:

https://en.wikipedia.org/wiki/Business_Process_Execution_Language

[582] “Apache ode.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_ODE

[583] “ODE website.” Web page [Online]. Available: <http://ode.apache.org/>

[584] Apache Software Foundation, “Omid project incubation status.” Web page, Apr-2016 [Online]. Available: <https://mrql.incubator.apache.org/>

[585] F. Perez-Sorrosal, O. Shacham, K. Tsioutsiouliklis, and E. Bortnikov, “Omid’s first step in the apache community.” Web page, Sep-2016 [Online]. Available: <https://yahooeng.tumblr.com/post/151015726181/omids-first-step-in-the-apache-community>

[586] Apache, “About oodt.” Web page [Online]. Available: <http://oodt.apache.org/>

[587] M. Islam et al., “Oozie: Towards a scalable workflow management system for hadoop,” in Proceedings of the 1st acm sigmod workshop on scalable workflow execution engines and technologies, 2012, p. 4.

[588] “Why Oozie? Just a simple Hadoop DBA.” Web page, 13-Feb-2017 [Online]. Available: <https://prodlife.wordpress.com/2013/12/09/why-oozie/>

[589] “Oozie - Apache Oozie Workflow Scheduler for Hadoop.” Web

page, 13-Feb-2017 [Online]. Available: <http://oozie.apache.org/>

[590] The Open MPI Project, “Open MPI: Open Source High Performance Computing.” Web page, Feb-2017 [Online]. Available: <https://www.open-mpi.org>

[591] E. Gabriel et al., “Open MPI: Goals, concept, and design of a next generation MPI implementation,” in Proceedings, 11th european pvm/mpi users’ group meeting, 2004, pp. 97-104 [Online]. Available: <https://www.open-mpi.org/papers/euro-pvmmpi-2004-overview/euro-pvmmpi-2004-overview.pdf>

[592] “OpenCV.” Web page [Online]. Available: <http://opencv.org/opencv-3-2.html>

[593] “About cv.” Webpage, 2018 [Online]. Available: <https://opencv.org/about.html>

[594] A. Dhaigude, “OpenCV-usecases.” Webpage, Oct-2016 [Online]. Available: <https://github.com/ad8454/OpenCV-UseCases>

[595] “OpenCV.” Webpage, Aug-2018 [Online]. Available: <https://en.wikipedia.org/wiki/OpenCV>

[596] Wikipedia, “OPeNDAP.” Web page, Jan-2005 [Online]. Available: <https://en.wikipedia.org/wiki/OPeNDAP>

[597] ferret, “OPeNDAP usage in ferret.” Web page [Online]. Available: <https://ferret.pmel.noaa.gov/Ferret/documentation/opendap/opendausage-in-ferret>

[598] jimg, “Welcome.” Web page, Sep-2018 [Online]. Available: <https://www.opendap.org/>

[599] jimg, “Comparing price and performance of three cloud-based data-storage architectures to optimize use of s3 in aws.” Web page, Dec-2018 [Online]. Available: https://www.opendap.org/index.php/about/workshops-and-presentations/S3_Optimization_AGU_2017

- [600] gdal, “DODS/opendap.” Web page [Online]. Available: https://www.gdal.org/drv_dods.html
- [601] “OpenID.” Web page [Online]. Available: <http://openid.net/>
- [602] “OpenID.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/OpenID>
- [603] oath, “User authentication with oauth 2.0,” 2014 [Online]. Available: <https://oauth.net/articles/authentication/>
- [604] M. Slot, “Beginner’s guide to openid phishing.” Web page, Nov-2008 [Online]. Available: <https://blog.rootshell.be/2008/11/05/beginners-guide-to-openid-phishing/>
- [605] Apache, “Apache openjpa.” Web page, Dec-2016 [Online]. Available: <http://openjpa.apache.org/>
- [606] K. Sutter, “Apache openjpa.” Web page, 2007 [Online]. Available: https://en.wikipedia.org/wiki/Apache_OpenJPA
- [607] OpenNebula, “OpenNebula concepts and terminology.” Web page, Oct-2018 [Online]. Available: http://docs.opennebula.org/5.6/intro_release_notes/concepts_terminology.html
- [608] R. Moreno-Vozmediano, Montero, and I. M. Llorente, “IaaS cloud architecture: From virtualized datacenters to federated cloud infrastructures,” Computer, vol. 45, no. 12, pp. 65–72, Dec. 2012.
- [609] “OpenPBS.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Portable_Batch_System
- [610] PBS, “OpenPBS.” Web page [Online]. Available: <http://www.pbspro.org/>
- [611] OpenRefine, “OpenRefine.” Web page, Feb-2016 [Online]. Available: <http://openrefine.org/>

- [612] “Heat - openstack.” Web page [Online]. Available: <https://wiki.openstack.org/wiki/Heat>
- [613] P. Julio Villarreal, “A quick introduction to openstack heat.” Web page, Jan-2017 [Online]. Available: <http://superuser.openstack.org/articles/quick-intro-openstack-heat/>
- [614] S. Lowe, “An introduction to openstack heat.” Web page, May-2014 [Online]. Available: <http://blog.scottlowe.org/2014/05/01/an-introduction-to-openstack-heat/>
- [615] D. V. D. Veen and L. Wrangler, “Introduction to Ironic.” Web page; OpenStack.org, Mar-2015 [Online]. Available: <https://docs.openstack.org/developer/ironic/deploy/user-guide.html>
- [616] Openstack, “Keystone, the openstack identity service.” Web page, 2018 [Online]. Available: <https://docs.openstack.org/keystone/latest/>
- [617] openstack, “OpenStack services.” Web page [Online]. Available: <https://www.openstack.org/software/project-navigator/openstack-components#openstack-services>
- [618] openstack, “Keystone architecture.” Web page, 2018 [Online]. Available: <https://docs.openstack.org/keystone/pike/getting-started/architecture.html>
- [619] “Software >> openstack open source cloud computing software.” Web page [Online]. Available: <https://www.openstack.org/software/>
- [620] “Foundation >> openstack open source cloud computing software.” Web page [Online]. Available: <https://www.openstack.org/foundation/>
- [621] “Apache license, version 2.0.” Web page [Online]. Available: <https://www.apache.org/licenses/LICENSE-2.0>
- [622] T. Binz et al., OpenTOSCA – a runtime for tosca-based cloud

applications. Springer Berlin Heidelberg, 2013, pp. 692–695 [Online]. Available: http://dx.doi.org/10.1007/978-3-642-45005-1_62

[623] “OpenVZ virtuozzo container.” Web page [Online]. Available: https://openvz.org/Main_Page

[624] P. Padala, X. Zhu, Z. Wang, S. Singhal, and K. G. Shin, “Performance Evaluation of Virtualization Technologies for Server Consolidation,” Tech. Rep., 2007 [Online]. Available: <http://www.hpl.hp.com/techreports/2007/HPL-2007-59R1.pdf>

[625] OpenVZ, “Features.” Web Page, 2011 [Online]. Available: https://wiki.openvz.org/index.php?title=Features&mobileaction=toggle_view_mobile

[626] Wikipedia, “OpenVZ.” Web page, 2018 [Online]. Available: <https://en.wikipedia.org/wiki/OpenVZ>

[627] OpenVZ, “OpenVZ guide.” Web page [Online]. Available: <https://download.openvz.org/doc/OpenVZ-Users-Guide.pdf>

[628] “Oracle labs ppx.” Web page [Online]. Available: https://docs.oracle.com/cd/E56133_01/2.3.1/index.html

[629] “Parallel graph analytics (ppx).” Web page [Online]. Available: <http://www.oracle.com/technetwork/oracle-labs/parallel-graph-analytics/overview/index.html>

[630] “Background.” Web page [Online]. Available: <https://orc.apache.org/docs/>

[631] IBM, “OSGi, the dynamic module system for java.” online, 1999 [Online]. Available: <https://www.osgi.org>

[632] Parasol, “Parasol.” Web page [Online]. Available: <https://parasol.tamu.edu/research.php>

[633] A. S. Foundation, “Apache parquet.” Web page [Online]. Available: <https://parquet.apache.org/documentation/latest>

- [634] Wikipedia, "Programming with big data in r." Web Page [Online]. Available:
https://en.wikipedia.org/wiki/Programming_with_Big_Data_in_R
- [635] "Pegasus." Web page [Online]. Available: <https://pegasus.isi.edu/>
- [636] "Pentaho." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Pentaho>
- [637] "Any analytics, any data, simplified." Web page [Online]. Available: <http://www.pentaho.com/product/product-overview>
- [638] PETSc, "PETSc documentation." [Online]. Available: <https://www.mcs.anl.gov/petsc/documentation/faq.html#computers>
- [639] Wikipedia, "Portable, extensible toolkit for scientific computation." [Online]. Available: https://en.wikipedia.org/wiki/Portable,_Extensible_Toolkit_for_Scientific_computation
- [640] A. N. Laboratory, "PETSc users manual." Sep-2018 [Online]. Available: <https://www.mcs.anl.gov/petsc/petsc-current/docs/manual.pdf>
- [641] J. Kestelyn, "Phoenix in 15 minutes or less." Web page, Mar-2013 [Online]. Available: <http://blog.cloudera.com/blog/2013/03/phoenix-in-15-minutes-or-less/>
- [642] Wikipedia, "Apache phoenix." Web page, Jan-2017 [Online]. Available: <https://en.m.wikipedia.org/wiki/Apache\Phoenix>
- [643] Apache Software Foundation, "Apache phoenix: OLTP and operational analytics for apache hadoop." Web page, Jan-2017 [Online]. Available: <http://phoenix.apache.org/>
- [644] A. Seligman, "Apache phoenix - a small step for big data." Web page, May-2014 [Online]. Available: <https://developer.salesforce.com/blogs/developer-relations/2014/05/apache-phoenix-small-step-big-data.html>

- [645] A. Avram, "Phoenix: Running sql queries on apache hbase [updated]." Web page, Jan-2013 [Online]. Available: <https://www.infoq.com/news/2013/01/Phoenix-HBase-SQL>
- [646] I. Szegedi, "Apache phoenix - an sql driver for hbase." Web page, May-2014 [Online]. Available: <https://bighadoop.wordpress.com/2014/05/17/apache-phoenix-an-sql-driver-for-hbase/>
- [647] R. M. Shakil Akhtar, Pro apache phoenix: An sql driver for hbase. Apress [Online]. Available: <https://books.google.com/books?id=EaTPDQAAQBAJ&printsec=frontcover&dq=aPACHE+phoenix&hl=en>
- [648] Apache Software Foundation, "Phoenix storage handler for apache hive." Web page, 2018 [Online]. Available: https://phoenix.apache.org/hive_storage_handler.html
- [649] Apache Software Foundation, "Apache pig integration." Web page, 2018 [Online]. Available: https://phoenix.apache.org/pig_integration.html
- [650] V. Harvey, "How to install apache phoenix," PhD thesis, 2017 [Online]. Available: <https://www.thesisscientist.com/docs/Others/2fc78636-eab1-4009-b2fa-5acb3fd94105>
- [651] I. Raicu, Y. Zhao, C. Dumitrescu, I. Foster, and M. Wilde, "Falkon: A fast and light-weight tasK executiON framework," in ACM sc07, 2007.
- [652] J.-S. Kim, S. Rho, S. Kim, S. Kim, S. Kim, and S. Hwang, "HTCaaS: Leveraging distributed supercomputing infrastructures for large-scale scientific computing," in ACM mtags 13, 2013.
- [653] M. Turilli, M. Santcroos, and S. Jha, "A comprehensive perspective on pilot-job systems," in ACM arXiv, 2016, pp. 1–26.
- [654] "About pivotal gemfire." Web page [Online]. Available:

http://gemfire.docs.pivotal.io/gemfire/getting_started/gemfire_overview.html

[655] S. Gollapudi, Getting started with greenplum for big data analytics. Packt Publishing, 2013 [Online]. Available: http://www.ebook.de/de/product/21653990/sunila_gollapudi_getting_started_with_greenplum_for_big_data_analytics.html

[656] Pivotal Software Inc, “Gpfdist.” Web page, 2017 [Online]. Available:

http://gpdb.docs.pivotal.io/4330/utility_guide/admin_utilities/gpfdist.html

[657] J. I. Cohen, L. Lonergan, and C. E. WeltonCohen, “Integrating map-reduce into a distributed relational database.” Google Patents, Dec-2016 [Online]. Available:

<https://www.google.com/patents/US9514188>

[658] “Pivotal greenplum: The open source massively parallel data warehouse.” Web page [Online]. Available: <https://pivotal.io/pivotal-greenplum>

[659] “Pivotal greenplum database.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Pivotal_Greenplum_Database

[660] Pivotal, “Pivotal.” Web page [Online]. Available: <https://run.pivotal.io/features/>

[661] E. Agullo, “Numerical linear algebra on emerging architectures: The magma and plasma projects,” Journal of Physics: Conference Series, 2009 [Online]. Available:

<http://iopscience.iop.org/article/10.1088/1742-6596/180/1/012037/pdf>

[662] “Plasma readme.” Web page, 2013 [Online]. Available: <http://icl.cs.utk.edu/projectsfiles/plasma/html/README.html>

[663] S. Tomov and J. Dongarra, “Matrix algebra on gpu and multicore architectures,” 2012 [Online]. Available: <https://www.olcf.ornl.gov/wp-content/training/electronic-structure-2012/ORNL-ESWorkshop.pdf>

[664] N. Patel, “PolyBase - how sql server does big data.” Web page

[Online]. Available: https://blogs.msdn.microsoft.com/premier_developer/2018/02/12/how-sql-server-does-big-data/

[665] “About postgresql.” Web page [Online]. Available: <https://www.postgresql.org/about/>

[666] The PostgreSQL Global Development Group, “PostgreSQL 9.5: UPSERT, row level security, and big data.” Web page [Online]. Available: <https://www.postgresql.org/about/news/1636/>

[667] “PostgreSQL history.” Web page [Online]. Available: <https://www.postgresql.org/about/history/>

[668] “Potree.” [Online]. Available: <http://potree.org/wp/about/>

[669] O. Martinez-Rubi et al., “Taming the beast: Free and open-source massive point cloud web visualization.” 2015 [Online]. Available: <http://repository.tudelft.nl/islandora/object/uuid:0472e0d1-ec75-465a-840e-fd53d427c177?collection=research>

[670] H. F. Halaoui, “A spatio temporal indexing structure for efficient retrieval and manipulation of discretely changing spatial data,” Journal of Spatial Science, vol. 53, no. 2, pp. 1–12, 2008 [Online]. Available: <http://dx.doi.org/10.1080/14498596.2008.9635146>

[671] G. Malewicz et al., “Pregel: A system for large-scale graph processing,” in Proceedings of the 2010 ACM SIGMOD international conference on management of data, 2010, pp. 135–146 [Online]. Available: <http://doi.acm.org/10.1145/1807167.1807184>

[672] Pregel, “Pregel: A system for large-scale graph processing.” May-2015 [Online]. Available: <https://blog.acolyer.org/2015/05/26/prege-a-system-for-large-scale-graph-processing/>

[673] “Presto.” Web page [Online]. Available: <https://prestodb.io/>

[674] Y. Chen et al., “A study of sql-on-hadoop systems,” in Big data benchmarks, performance optimization, and emerging hardware: 4th

and 5th workshops, bpoe 2014, salt lake city, usa, march 1, 2014 and hangzhou, china, september 5, 2014, revised selected papers, Springer International Publishing, 2014, pp. 154–166.

[675] “Protocol buffer.” Web page, Sep-2016 [Online]. Available: <https://developers.google.com/protocol-buffers/>

[676] M. Bernstein, “5 reasons to use protocol buffers instead of json for your next service.” Web page, Jun-2014 [Online]. Available: <https://codeclimate.com/blog/choose-protocol-buffers/>

[677] Google Developers, “Other languages.” Web page, Jul-2016 [Online]. Available: <https://developers.google.com/protocol-buffers/docs/reference/other>

[678] Google Developers, “Developer guide.” Web page, Aug-2018 [Online]. Available: <https://developers.google.com/protocol-buffers/docs/overview>

[679] Y. Shinde, “Protobuf performance comparison and points to consider when deciding if it’s right for you.” Web page, Apr-2016 [Online]. Available: <https://dzone.com/articles/protobuf-performance-comparison-and-points-to-make>

[680] Google Developers, “Techniques.” Web page, Sep-2018 [Online]. Available: <https://developers.google.com/protocol-buffers/docs/techniques>

[681] V. J. Setty, “Publish/subscribe for large-scale social interaction: Design, analysis and resource provisioning,” PhD thesis, Department of Informatics,UNIVERSITY OF OSLO, Faculty of Mathematics; Natural Sciences, University of Oslo, 2015 [Online]. Available: <https://www.duo.uio.no/bitstream/handle/10852/43117/1595-Setty-DUO-Thesis.pdf?sequence=1&isAllowed=y>

[682] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, “The many faces of publish/subscribe,” ACM Comput. Surv., vol. 35, no. 2, pp. 114–131, Jun. 2003 [Online]. Available:

<http://doi.acm.org/10.1145/857076.857078>

[683] C. Esposito, M. Ficco, F. Palmieri, and A. Castiglione, "A knowledge-based platform for big data analytics based on publish/subscribe services and stream processing," *Know.-Based Syst.*, vol. 79, no. C, pp. 3–17, May 2015 [Online]. Available: <http://dx.doi.org/10.1016/j.knosys.2014.05.003>

[684] Puppet Lab, "Puppet." Web page [Online]. Available: <https://puppet.com/>

[685] "Puppet software." Web page [Online]. Available: [https://en.wikipedia.org/wiki/Puppet_\(software\)](https://en.wikipedia.org/wiki/Puppet_(software))

[686] C. Nuggets, "How puppet works." Apr-2016 [Online]. Available: <https://www.youtube.com/watch?v=lxJQX2ipliY>

[687] Puppet, "Overview of puppet's architecture." [Online]. Available: <https://puppet.com/docs/puppet/4.6/architecture.html>

[688] T. Schaul et al., "PyBrain," *Journal of Machine Learning Research*, vol. 11, pp. 743–746, Mar. 2010 [Online]. Available: <http://dl.acm.org/citation.cfm?id=1756006.1756030>

[689] Z. Z, "PyBrain - a simple neural networks library in python." 2014 [Online]. Available: <http://fastml.com/pybrain-a-simple-neural-networks-library-in-python/>

[690] Wikipedia, "QEMU." Web page, Feb-2017 [Online]. Available: <https://en.wikipedia.org/wiki/QEMU>

[691] Qemu, "QEMU." Web page, Feb-2017 [Online]. Available: http://wiki.qemu-project.org/index.php/Main_Page

[692] R, "R programming language." Web Page, 18-Oct-2018 [Online]. Available: <https://www.r-project.org/>

[693] Wikipedia, "R(programming language)." Web Page, 18-Oct-2018 [Online]. Available:

[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

[694] D. Kopf, "If you want to upgrade your data analysis skills, which programming language should you learn?" Web Page, 2017 [Online]. Available: <https://qz.com/1063071/the-great-r-versus-python-for-data-science-debate/>

[695] "RabbitMQ, components." Web page [Online]. Available: <https://www.rabbitmq.com/>

[696] Y. Trudeau, "Exploring message brokers: RabbitMQ, kafka, activemq, and kestrel," Integration Zone, 2014 [Online]. Available: <https://dzone.com/articles/exploring-message-brokers>

[697] rasdaman GmbH, "Rasdaman." 2018 [Online]. Available: <https://en.wikipedia.org/wiki/Rasdaman>

[698] rasdaman GmbH, "Rasdaman." 2018 [Online]. Available: <http://tutorial.rasdaman.org/>

[699] rasdaman GmbH, "Rasdaman." 2018 [Online]. Available: <http://www.rasdaman.org/>

[700] Puppet Labs, "Home.PuppetLabs/razor wiki." Web page [Online]. Available: <https://github.com/puppetlabs/Razor/wiki>

[701] Puppet Labs, "Introducing razor, a next generation provisioning solution-puppet." Web page [Online]. Available: <https://puppet.com/blog/introducing-razor-a-next-generation-provisioning-solution>

[702] "RCFileCat - apache hive." Web page [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/RCFileCat>

[703] "RCFile cat." Web page [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/RCFileCat>

[704] Y. He et al., "RCFile: A fast and space-efficient data placement structure in mapreduce-based warehouse systems," in Data

engineering (icde), 2011 ieee 27th international conference on engineering, 2011, pp. 1199–1208.

[705] Red Hat, Inc., “paas-openshift components.” Web page [Online]. Available: <https://blog.openshift.com/announcing-openshift-origin-the-open-source-platform-as-a-service-paas/>

[706] Red Hat, Inc., “developers-openshift components.” Web page [Online]. Available: <https://developers.openshift.com/>

[707] Red Hat, Inc., “openshift components.” Web page [Online]. Available: <https://www.openshift.org/>

[708] Red Hat, Inc., “openshift-blog components.” Web page [Online]. Available: <https://blog.openshift.com/getting-started-with-openshift-origin-the-open-source-platform-as-a-service-paas/>

[709] “Introduction to radix.” Web page [Online]. Available: <https://redis.io/topics/introduction>

[710] B. Kepes, “Redis labs hires the creator of redis, salvatore sanfilippo,” Network World, 2015 [Online]. Available: <https://www.networkworld.com/article/2947895/opensource-subnet/redis-labs-hires-the-creator-of-redis-salvatore-sanfilippo.html>.

[711] Byung-Gon Chun, “REEFProposal - incubator.” Web page, Aug-2014 [Online]. Available: <https://wiki.apache.org/incubator/ReefProposal>

[712] RIAK KV, “RIAK kv.” Web Page [Online]. Available: <http://basho.com/products/riak-kv/>

[713] R. Eck, “An overview of riak: An open source nosql database.” Web Page [Online]. Available: <http://www.monitis.com/blog/an-overview-of-riak-an-open-source-nosql-database/>

[714] CoreOS, “Rkt overview.” Web page, 2018 [Online]. Available: <https://coreos.com/rkt/>

[715] CoreOS, “Rkt architecture.” Web page, 2018 [Online]. Available: <https://coreos.com/rkt/docs/latest-devel/architecture.html>

[716] A. de Jonge, “Moving from docker to rkt.” Web page, Jul-2016 [Online]. Available: <https://medium.com/@adriaandejonge/moving-from-docker-to-rkt-310dc9aec938>

[717] Open Source Robotics Foundation, “About ros.” Web page, Mar-2017 [Online]. Available: <http://www.ros.org/about-ros/>

[718] National Instruments, “A layered approach to designing robot software.” Web page, Mar-2017 [Online]. Available: <http://www.ni.com/white-paper/13929/en/>

[719] “Rocks.” Web page, 2017 [Online]. Available: <https://www.rockscluster.org>

[720] R. Punnoose, A. Crainiceanu, and D. Rapp, “Rya: A scalable rdf triple store for the clouds,” in Proceedings of the 1st international workshop on cloud intelligence, 2012, pp. 4:1–4:8 [Online]. Available: <http://doi.acm.org/10.1145/2347673.2347677>

[721] R. W. Group, “Resource description framework (rdf).” Web page, Feb-2014 [Online]. Available: <https://www.w3.org/RDF/>

[722] Apache, “Apache rya.” Web page [Online]. Available: <https://rya.apache.org/>

[723] “S4: Distributed stream computing platform.” Web page [Online]. Available: <http://incubator.apache.org/s4/>

[724] “S4: S4 0.6.0 overview.” Web page [Online]. Available: <http://incubator.apache.org/s4/doc/0.6.0/overview>

[725] M. Matthieu, “S4 0.5.0 overview.” Web page, Feb-2013 [Online]. Available: <https://cwiki.apache.org/confluence/display/S4/S4+0.5.0+Overview>

[726] S. Jha, H. Kaiser, A. Merzky, and O. Weidner, “Grid

interoperability at the application level using saga," in 07 proceedings of the third ieee international conference on e-science and grid computing, 2007.

[727] T. Goodale et al., "A simple api for grid applications." Web page, Sep-2013 [Online]. Available: <https://www.ogf.org/documents/GFD.90.pdf>

[728] "OpenStack." Web page [Online]. Available: <https://www.openstack.org/>

[729] "Sahara." Web page [Online]. Available: <http://docs.openstack.org/developer/sahara/>

[730] Wikipedia, "Salt." Web page, Oct-2018 [Online]. Available: [https://en.wikipedia.org/wiki/Salt_\(software\)](https://en.wikipedia.org/wiki/Salt_(software))

[731] Sebastian Braun, "Introduction to salt and saltstack." Web page, Feb-2017 [Online]. Available: <https://www.mirantis.com/blog/introduction-to-salt-and-saltstack/>

[732] SaltStack, "INTRODUCTION to salt." Web page, Oct-2018 [Online]. Available: <https://docs.saltstack.com/en/latest/topics/>

[733] A. Sandil, "SSO strategy: Authentication (saml) -vs- authorization (oauth)." Web page [Online]. Available: <https://www.linkedin.com/pulse/sso-strategy-authentication-vs-authorization-saml-oauth-sandil>

[734] "Apache Samza," Wikipedia. Web page, 13-Feb-2017 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Apache_Samza&oldid=764035647

[735] "Samza." Web page, 13-Feb-2017 [Online]. Available: <http://samza.apache.org/>

[736] "Apache Samza, LinkedIn's Framework for Stream Processing," The New Stack. Web page, 13-Feb-2017 [Online]. Available: <https://thenewstack.io/apache-samza-linkedin-s-framework-for->

stream-processing/

[737] SAP, Web page [Online]. Available: <https://www.sap.com/products/hana.html>

[738] Wikipedia, Web page [Online]. Available: https://en.wikipedia.org/wiki/SAP_HANA

[739] C. Gaska, "What is sap hana? [2018 updated guide]." Web page, Jun-2018 [Online]. Available: <https://symmetrycorp.com/blog/what-is-sap-hana/>

[740] SAP, Web page [Online]. Available: <https://www.sap.com/products/hana/features.html/>

[741] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan, "Interpreting the data: Parallel analysis with sawzall," Scientific Programming Journal, vol. 13, no. 4, pp. 277–298, Oct. 2005 [Online]. Available: <https://research.google.com/archive/sawzall-sciprog.pdf>

[742] R. Rosario, "Exciting tools for big data: S4, sawzall and mrjob!" Web page, Nov-2010 [Online]. Available: <http://www.bytемining.com/2010/11/exciting-tools-for-big-data-s4-sawzall-and-mrjob/>

[743] Alphabet, Inc., "Szl - overview.wiki." Code Repository [Online]. Available: <https://code.google.com/archive/p/szl/wikis/Overview.wiki>

[744] netlib, "ScalAPACK users guide." Web page [Online]. Available: <http://www.netlib.org/scalapack/slug/>

[745] M. Stonebraker, "SciDB: An open-source dbms for scientific data," ERCIM News, no. 89, p. 56, Apr. 2012 [Online]. Available: <https://ercim-news.ercim.eu/en89/special/scidb-an-open-source-dbms-for-scientific-data>

[746] J. Brownlee, "A gentle introduction to scikit-learn: A python machine learning library." Web page, 16-Apr-2014 [Online]. Available: <http://machinelearningmastery.com/a-gentle-introduction-to-scikit->

[learn-a-python-machine-learning-library/](#)

[747] “Scikit learn.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/Scikit-learn>

[748] “Hadoop data security and sentry.” Web Page [Online]. Available: <https://www.ibm.com/developerworks/library/se-hadoop/index.html>

[749] Aduna, “sesame components.” Web page [Online]. Available: <https://projects.eclipse.org/projects/technology.rdf4j>

[750] A. L. V. Jianbin Fang and H. Sips, “Sesame: A user-transparent optimizing framework for many-core processors,” in IEEE/acm international symposium on cluster, cloud, and grid computing, 2013, pp. 70–73.

[751] WikiStart, “Welcome to the son of grid engine project.” Web page [Online]. Available: <https://arc.liv.ac.uk/trac/SGE>

[752] Univa, “Welcome to the son of grid engine project.” Web page [Online]. Available: <http://www.univa.com/products/>

[753] Wikipedia, “Univa grid engine.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Univa_Grid_Engine

[754] Wikipedia, “Oracle grid engine - wikipedia.” Web page, Feb-2017 [Online]. Available: https://en.wikipedia.org/wiki/Oracle_Grid_Engine

[755] Softparorama, “Grid engine as a high quality unix/linux batch system.” Web page [Online]. Available: http://www.softpanorama.org/HPC/Grid_engine/index.shtml

[756] Bioinformatics, “Sun grid engine for beginners.” Web page [Online]. Available: http://bioinformatics.mdc-berlin.de/intro2UnixandSGE/sun_grid_engine_for_beginners/README

[757] Rayson, “Running a 10,000-node grid engine cluster in amazon ec2.” Web page, Nov-2012 [Online]. Available:

<http://blogs.scalablelogic.com/2012/11/running-10000-node-grid-engine-cluster.html>

[758] Engle et al., "Shark: Fast data analysis using coarse-grained distributed memory," in Proceedings of the 2012 ACM SIGMOD international conference on management of data, 2012, pp. 689–692 [Online]. Available: <http://doi.acm.org/10.1145/2213836.2213934>

[759] R. Xin, J. Rosen, M. Zaharia, M. Franklin, S. Shenker, and I. Stoica, "Shark: SQL and rich analytics at scale," ACM SIGMOD Conference, 2013 [Online]. Available: <https://amplab.cs.berkeley.edu/publication/shark-sql-and-rich-analytics-at-scale/>

[760] R. Xin, "Shark, spark sql, hive on spark, and the future of sql on apache spark." Web page, 2014 [Online]. Available: <https://databricks.com/blog/2014/07/01/shark-spark-sql-hive-on-spark-and-the-future-of-sql-on-spark.html>

[761] slideshare, "Spark and shark." Web page, 2012 [Online]. Available: https://www.slideshare.net/Hadoop_Summit/spark-and-shark

[762] B. Lorica, "Shark: Real-time queries and analytics for big data," 2012 [Online]. Available: <https://www.oreilly.com/ideas/shark-real-time-queries-and-analytics-for-big-data>

[763] "SLURM." Web page [Online]. Available: <https://slurm.schedmd.com/>

[764] Wikipedia, "Slurm workload manager." Web page [Online]. Available: https://en.wikipedia.org/wiki/Slurm_Workload_Manager

[765] Nvidia Developer, Web page [Online]. Available: <https://developer.nvidia.com/slurm>

[766] SchedMd, "Slurm website." Web page, Mar-2013 [Online]. Available: <https://slurm.schedmd.com/overview.html>

[767] M. T. Jones, “Optimizing resource management in supercomputers with slurm.” Web page, May-2012 [Online]. Available: <https://www.ibm.com/developerworks/library/l-slurm-utility/index.html>

[768] Cisco, “Snort.” Web Page, 18-Oct-2018 [Online]. Available: <https://www.snort.org/>

[769] Wikipedia, “Snort (software).” Web Page, 18-Oct-2018 [Online]. Available: [https://en.wikipedia.org/wiki/Snort_\(software\)](https://en.wikipedia.org/wiki/Snort_(software))

[770] Cisco, “Snort: The world’s most widely deployed ips technology.” Web Page, 2014 [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/security/brief_c17-733286.html

[771] M. P. Brennan, “Using snort for a distributed intrusion detection system,” SANS Institute, 2002 [Online]. Available: <https://www.sans.org/reading-room/whitepapers/detection/snort-distributed-intrusion-detection-system-352>

[772] Jake Luciani, “Solandra wiki.” Code Repository, Feb-2017 [Online]. Available: <https://github.com/tjake/Solandra/wiki/Solandra-Wiki>

[773] The Apache Software Foundation, “Spark sql.” Web page, Feb-2016 [Online]. Available: <http://spark.apache.org/sql/>

[774] ACM, Inc., “Spark sql: Relational data processing in spark.” Web page, Feb-2016 [Online]. Available: <http://dl.acm.org/citation.cfm?id=2742797>

[775] S. Penchikala, “Big data processing with apache spark part 3: Spark streaming.” Web page, Jan-2016 [Online]. Available: <https://www.infoq.com/articles/apache-spark-streaming>

[776] “Taming big data with spark streaming for real time data processing.” Web page, Mar-2017 [Online]. Available:

<https://www.dezyre.com/article/taming-big-data-with-spark-streaming-for-real-time-data-processing/337>

[777] A. G. Shoro and T. R. Soomro, "Big data analysis: Apache spark perspective," Global Journal of Computer Science and Technology, vol. 15, no. 1, 2015 [Online]. Available: <http://www.computerresearch.org/index.php/computer/article/view/1>

[778] Splunk, Inc., "Splunk enterprise overview." Web page, Feb-2015 [Online]. Available: <http://docs.splunk.com/Documentation/Splunk/6.5.2/Overview/About>

[779] Wikipedia, "Microsoft sql server." Web Page [Online]. Available: https://en.wikipedia.org/wiki/Microsoft_SQL_Server

[780] Microsoft, "Editions and supported features of sql server 2016." Web Page [Online]. Available: <https://docs.microsoft.com/en-us/sql/sql-server/editions-and-components-of-sql-server-2016?view=sql-server-2017>

[781] SQLite, "About sqlite." Web page [Online]. Available: <https://www.sqlite.org/about.html>

[782] J. C. E. Margaret Rouse, "ACID (atomicity, consistency, isolation, and durability)." Web page, Jul-2006 [Online]. Available: <http://searchsqlserver.techtarget.com/definition/ACID>

[783] tutorialspoint, "SQLite overview." Web page [Online]. Available: https://www.tutorialspoint.com/sqlite/sqlite_overview.htm

[784] SQLite, "Appropriate uses for sqlite." Web page [Online]. Available: <https://www.sqlite.org/whentouse.html>

[785] Wikipedia, "SQLite." Web page [Online]. Available: <https://en.wikipedia.org/wiki/SQLite>

[786] "SQLite as an application file format." Web page [Online]. Available: https://www.sqlite.org/aff_short.html

[787] “SQLite download page.” Web page [Online]. Available: <https://www.sqlite.org/download.html>

[788] The Apache Software Foundation, “Sqoop.” Web page [Online]. Available: <http://sqoop.apache.org/>

[789] Wikipedia, “Sqoop.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/Sqoop>

[790] “SSH - wikipedia.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Secure_Shell

[791] “OpenSSH - wikipedia.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/OpenSSH>

[792] HPE, “HPE helion stackato.” Web page [Online]. Available: <https://www.hpe.com/us/en/software/multi-cloud-platform.html>

[793] M. Kavis, “THE pros and cons of private and public paas.” Web page, Jun-2013 [Online]. Available: <https://www.virtualizationpractice.com/the-pros-and-cons-of-private-and-public-paas-21961/>

[794] M. H. Iqbal and T. R. Soomro, “Big data analysis: Apache storm perspective,” in International journal of computer trends and technology (ijctt)-2015, 2015.

[795] “Apache storm - concepts.” Web page [Online]. Available: <http://storm.apache.org/releases/1.0.2/Concepts.html>

[796] Wikipedia, “Apache flink.” Web page, Feb-2017 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Flink

[797] O. B. A. I. O. A. S. R. A. A. Singhal, “Summingbird.” Web page [Online]. Available: <https://github.com/twitter/summingbird>

[798] O. Boykin, S. Ritchie, I. O’Connell, and J. Lin, “Summingbird: A framework for integrating batch and online mapreduce computations,” Proceedings of the VLDB Endowment, vol. 7, no. 13,

pp. 1441–1451, 2014.

[799] Wikipedia, “Swift (programming language).” Web page, Feb-2017 [Online]. Available:

[https://en.wikipedia.org/wiki/Swift_\(programming_language\)](https://en.wikipedia.org/wiki/Swift_(programming_language))

[800] “Tableau tutorial.” Web page [Online]. Available: <https://casci.umd.edu/wp-content/uploads/2013/12/Tableau-Tutorial.pdf>

[801] “Tableau technology.” Web page [Online]. Available: <https://www.tableau.com/products/technology>

[802] ApacheTajo, “Apache tajo: A big data warehouse system on hadoop.” Web page, Oct-2018 [Online]. Available: <http://tajo.apache.org/>

[803] Tutorialspoint, “Apache tajo quick guide.” Web page, Oct-2018 [Online]. Available: https://www.tutorialspoint.com/apache_tajo/apache_tajo_quick_guide

[804] Apache, “Apache taverna.” Web Page, 04-Nov-2018 [Online]. Available: <https://taverna.incubator.apache.org/>

[805] Wikipedia, “Apache taverna.” Web Page, 18-Oct-2018 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Taverna

[806] Apache, “Introduction-what is taverna.” Web Page, 18-Oct-2018 [Online]. Available: <https://taverna.incubator.apache.org/introduction/>

[807] PredictiveAnalyticsToday, “Apache taverna-review.” Web Page, 18-Oct-2018 [Online]. Available: <https://www.predictiveanalyticstoday.com/apache-taverna/>

[808] K. Wolstencroft et al., “The taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud,” Nucleic Acids Research, vol. 41, nos. W557 – W561, 2013 [Online]. Available:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692062/>

[809] “TensorFlow.” Web page [Online]. Available: <https://www.tensorflow.org>

[810] S. Yegulalp, “What is tensorflow? The machine learning library explained.” Jun-2018 [Online]. Available: <https://www.infoworld.com/article/3278008/tensorflow/what-is-tensorflow-the-machine-learning-library-explained.html>

[811] M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems.” 2015 [Online]. Available: <https://www.tensorflow.org/>

[812] “TensorFlow graphs and sessions.” Webpage, Sep-2018 [Online]. Available: <https://www.tensorflow.org/guide/graphs>

[813] “TensorFlow: Machine learning for everyone.” Youtube Web, Feb-2017 [Online]. Available: <https://www.youtube.com/watch?v=mWI45NkFBOc&vl=en>

[814] T. Pott, “What on earth is terraform: Life support for explorers of terrifying alien worlds.” Web page, Dec-2017 [Online]. Available: https://www.theregister.co.uk/2017/12/06/what_is_terraform/

[815] “Terraform.” Web page [Online]. Available: <https://www.terraform.io/intro/index.html>

[816] “Terraform.” Web page [Online]. Available: <https://stackshare.io/terraform/in-stacks>

[817] “Tez.” Web page, 2017 [Online]. Available: <https://hortonworks.com/apache/tez>

[818] “Theano.” Web page [Online]. Available: <http://deeplearning.net/software/theano/introduction.html>

[819] “Three.js.” Web page, Mar-2017 [Online]. Available: <https://en.wikipedia.org/wiki/Three.js>

- [820] “Creating a scene.” Web page, Dec-2016 [Online]. Available: <https://threejs.org/docs/index.html#Manual/Getting Started/Creating>
- [821] Apache, “About thrift.” Web page [Online]. Available: <https://thrift.apache.org/>
- [822] K. Maeda, “Performance evaluation of object serialization libraries in xml, json and binary formats,” in Digital information and communication technology and it’s applications (dictap), 2012 second international conference on, 2012, pp. 177-182.
- [823] M. Slee, A. Agarwal, and M. Kwiatkowski, “Thrift: Scalable cross-language services implementation,” Thrift, Apr. 2007 [Online]. Available: <https://thrift.apache.org/static/files/thrift-20070401.pdf>
- [824] “Apache tika.” Web page [Online]. Available: <https://tika.apache.org/>
- [825] ApacheTinkerPopDoc, “Apache tinkerpop.” Web page, 2016 [Online]. Available: <http://tinkerpop.apache.org/docs/3.1.1-incubating/tutorials/getting-started/>
- [826] Dylan Raithel, “Apache tinkerpop graduates to top-level project.” Web page, 2016 [Online]. Available: <https://www.infoq.com/news/2016/06/tinkerpop-top-level-apache/>
- [827] Apache Tinker Pop Home, “Apache tinkerpop home.” Web page, 2016 [Online]. Available: <https://tinkerpop.apache.org/>
- [828] Titan, “Titan documentation.” Web page [Online]. Available: <http://s3.thinkaurelius.com/docs/titan/1.0.0/arch-overview.html>
- [829] Wikipedia, “Torch (machine learning).” 2014 [Online]. Available: [https://en.wikipedia.org/wiki/Torch_\(machine_learning\)](https://en.wikipedia.org/wiki/Torch_(machine_learning))
- [830] Netsyslab, “TOTEM.” Web page [Online]. Available: <http://netsyslab.ece.ubc.ca/wiki/index.php/Totem>
- [831] Cardiff University, “Triana documentation.” Cardiff University;

Web page, 2012 [Online]. Available: <http://www.trianacode.org/>

[832] T. A. S. Foundation, “Trident tutorial.” Web page [Online]. Available: <http://storm.apache.org/releases/0.10.1/Trident-tutorial.html>

[833] T. A. S. Foundation, “Trident api overview.” Web page [Online]. Available: <http://storm.apache.org/releases/1.0.0/Trident-API-Overview.html>

[834] “Twister.” Dictionary [Online]. Available: <http://backstopmedia.booktype.pro/big-data-dictionary/twister/>

[835] “Iterative map reduce.” Web page [Online]. Available: <https://iterativemapreduce.weebly.com/twister.html>

[836] Twitter, “Heron.” Web page, Nov-2018 [Online]. Available: <https://apache.github.io/incubator-heron/>

[837] K. Ramasamy, “Flying faster with twitter heron.” Web page, Jun-2015 [Online]. Available: https://blog.twitter.com/engineering/en_us/a/2015/flying-faster-with-twitter-heron.html

[838] K. Ramasamy, “Introduction to heron.” Web page, Aug-2017 [Online]. Available: <https://streamli.io/blog/intro-to-heron>

[839] FAL Labs, “Kyoto tycoon: A handy cache/storage server.” Web page [Online]. Available: <http://fallabs.com/kyototycoon/>

[840] “Tyrant blog.” Web page [Online]. Available: https://www.percona.com/blog/2009/10/19/mysql_memcached_tyrant

[841] “Tyrant fallabs.” Web page [Online]. Available: <http://fallabs.com/tokyotyrant/>

[842] “Kyoto tycoon.” Web page [Online]. Available: <http://fallabs.com/kyototycoon/>

[843] Wikipedia, "Tokyo cabinet and kyoto cabinet." Web page [Online]. Available:

https://en.wikipedia.org/wiki/Tokyo_Cabinet_and_Kyoto_Cabinet

[844] "Tyrant fallabs." Web page [Online]. Available: <http://fallabs.com/tokyotyrant/spex.html>

[845] Redstation, "Bare metal cloud vs. IaaS - are they the same thing?" Web page [Online]. Available: <http://www.redstation.com/blog/bare-metal-cloud-vs.-iaas-are-they-the-same-thing/>

[846] J. Castro, "What is maas exactly?" Web page, Jun-2014 [Online]. Available: <https://askubuntu.com/questions/486701/what-is-maas-exactly>

[847] MaaS.io, Web page [Online]. Available: <https://maas.io/>

[848] MaaS.io, "How if works." Web page [Online]. Available: <https://maas.io/how-it-works>

[849] S. Merrill, "Canonical Metal-as-a-Service: Not Quite As Cool As It Sounds," Tech Crunch, 2011 [Online]. Available: <https://techcrunch.com/2012/04/04/canonical-metal-as-a-service-not-quite-as-cool-as-it-sounds/>

[850] Wikipedia, "UIMA." Web page, May-2018 [Online]. Available: <https://en.wikipedia.org/wiki/UIMA>

[851] "Apache uima." Web page [Online]. Available: <https://uima.apache.org/>

[852] "Unstructured information management architecture sdk." Web page [Online]. Available: <https://www.ibm.com/developerworks/data/downloads/uima/index.html>

[853] K. Shaw, "What is a hypervisor?" Web page, Dec-2017.

[854] K. Purdy, "VirtualBox 3.2 beta virtualizes mac os x (on macs)."

Web page, May-2010.

[855] M. Schroeder and J. Saltzer, "A hardware architecture for implementing protection rings." Web page, Mar-1972.

[856] B. Roussey, "Here's why virtualbox is a fantastic virtualization management tool." Jan-2017.

[857] Wikipedia, "/vmware esxi." Web page [Online]. Available: https://en.wikipedia.org/wiki/VMware_ESXi

[858] VMware, "VMWARE." Web page [Online]. Available: <http://www.vmware.com/products/esxi-and-esx.html>

[859] LinkedIn, "Voldemort." 2018 [Online]. Available: [https://en.wikipedia.org/wiki/Voldemort_\(distributed_data_store\)](https://en.wikipedia.org/wiki/Voldemort_(distributed_data_store))

[860] LinkedIn, "Voldemort." 2018 [Online]. Available: <https://github.com/voldemort/voldemort>

[861] LinkedIn, "Voldemort," Solving Big Data Challenges for Enterprise Application Performance Management, 2018 [Online]. Available: <https://github.com/voldemort/voldemort>

[862] VoltDB, "VoltDB." Web page [Online]. Available: <https://www.voltdb.com/>

[863] vmware, "VCloud." Web page [Online]. Available: <http://www.vmware.com/products/vcloud-suite.html>

[864] Bipin, "Difference between vSphere, esxi and vCenter." Web page, Aug-2012 [Online]. Available: <http://www.mustbegeek.com/difference-between-vsphere-esxi-and-vcenter/>

[865] "Whirr technology." Web page [Online]. Available: <https://whirr.apache.org/>

[866] "Whirr technology." Web page [Online]. Available:

<http://www.slideshare.net/huguk/apache-whirr>

[867] O. Kopp, "Eclipse winery user guide." Web page, 2018 [Online]. Available: <https://projects.eclipse.org/projects/soa.winery>

[868] Eclipse, "Eclipse winery user guide." Web page, Jan-2018 [Online]. Available: <http://eclipse.github.io/winery/user/>

[869] A. S. Foundation, "Apache wink." Web page [Online]. Available: <https://svn.apache.org/repos/infra/websites/production/wink/content>

[870] Wikipedia, "Java api for restful web services." Web page [Online]. Available:

https://en.wikipedia.org/wiki/Java_API_for_RESTful_Web_Services

[871] "Apache wink developer guide." Web page, Apr-2010 [Online]. Available:

<https://svn.apache.org/repos/infra/websites/production/wink/content>

[872] A. S. Foundation, Web page [Online]. Available: <http://attic.apache.org/projects/wink.html>

[873] J. Paul, "7 reasons to use spring mvc for developing restful web services in java." Web page, Feb-2018 [Online]. Available: <https://dzone.com/articles/7-reasons-to-use-spring-mvc-for-developing-restful>

[874] xcat, "Extreme cloud/cluster administration toolkit." Web page, 2015 [Online]. Available: <http://xcat-docs.readthedocs.io/en/stable/>

[875] ibm, "Extreme cloud administration toolkit." Web page [Online]. Available: <http://www-03.ibm.com/systems/technicalcomputing/xcat/>

[876] "Xen - wikipedia." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Xen>

[877] "Xen project overview." Web page [Online]. Available: https://wiki.xenproject.org/wiki/Xen_Project_Software_Overview

- [878] “Xen feature list.” Web page [Online]. Available: https://wiki.xenproject.org/wiki/Xen_Project_4.7_Feature_List
- [879] Cray Inc, “Cray yarcdata urika appliance.” Web page, 2017 [Online]. Available: <http://www.cray.com/products/Urika.aspx>
- [880] Cray Inc, “Cray urika-gd technical specification,” Cray Inc, technical report, 2014 [Online]. Available: <http://www.cray.com/sites/default/files/resources/Urika-GD-TechSpecs.pdf>
- [881] B. Robin and J. Rebecca, “THE nature of graph data,” The Bloor Group; The Bloor Group, Austin, TX 78720|512-524-3689, 2017 [Online]. Available: <http://web.cray.com/bloor-graph-data>
- [882] L. Sangkeun, R. S. Sreenivas, and L. Seung-Hwan, “Graph mining meets the semantic web.” in 2015 31st ieee international conference on data engineering workshops, Seoul, South Korea, 2015, pp. 53-58 [Online]. Available: https://www.researchgate.net/profile/Sreenivas_Rangan_Sukumar/pu
- [883] M. Rouse, “Introduction: Apache hadoop yarn.” Mar-2018 [Online]. Available: <https://searchdatamanagement.techtarget.com/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator>
- [884] T. A. S. Fondation, “Apache hadoop yarn - overview.” [Online]. Available: https://hortonworks.com/apache/yarn/#section_2
- [885] T. A. S. Fondation, “Apache hadoop yarn.” Apr-2018 [Online]. Available: <https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>
- [886] A. Zeppelin, “Apache zeppelin 0.7.0 documentation.” Web page; zeppelin.apache.org, Feb-2017 [Online]. Available: <https://zeppelin.apache.org/docs/0.7.0/>
- [887] O. Tezer, “How to work with the zeromq messaging library.”

Web page, Dec-2013 [Online]. Available:
<https://www.digitalocean.com/community/tutorials/how-to-work-with-the-zeromq-messaging-library>

[888] H. Powell, "A quick and dirty introduction to zeromq." Web page, Mar-2015 [Online]. Available:
<https://blog.scottlogic.com/2015/03/20/ZeroMQ-Quick-Intro.html>

[889] M. Sustrik, "ZeroMQ." Web page, 2017 [Online]. Available:
<https://www.aosabook.org/en/zeromq.html>

[890] P. Hintjens, "A quick and dirty introduction to zeromq." Web page, 2016 [Online]. Available: <http://zguide.zeromq.org/page:all>

[891] I. I. of Technology Department of Computer Science, "ZHT: A zero-hop distributed hashtable." Web page [Online]. Available: <http://datasys.cs.iit.edu/projects/ZHT/>

[892] B. Wiley, "Distributed hash ttable, part 1," Linux Journal, no. 114, Oct. 2003 [Online]. Available:
<http://www.linuxjournal.com/article/6797?page=0,0>

[893] T. Li et al., "ZHT: A light-weight reliable persistent dynamic scalable zero-hop distributed hash table," in 2013 ieee 27th international symposium on parallel and distributed processing, 2013, pp. 775–787 [Online]. Available: http://datasys.cs.iit.edu/publications/2013_IPDPS13_ZHT.pdf

[894] "Zookeeper - overview." Web page [Online]. Available:
<https://zookeeper.apache.org/doc/trunk/zookeeperOver.html>

[895] "Zookeeper - wikipedia." Web page [Online]. Available:
https://en.wikipedia.org/wiki/Apache_ZooKeeper

[896] "IBM - what is zookeeper." Web page [Online]. Available:
<http://www-01.ibm.com/software/data/infosphere/hadoop/zookeeper/>

[897] A. Group, "Alibaba cloud." Web page [Online]. Available:

[https://www.alibabacloud.com/about?
spm=a3c0i.7911826.677923.38.4419737ba9UIy6](https://www.alibabacloud.com/about?spm=a3c0i.7911826.677923.38.4419737ba9UIy6)

[898] “Alluxio open source,” Alluxio. [Online]. Available: <https://www.alluxio.org/docs/1.7/en/index.html>

[899] A. W. Services, “AWS api gateway.” 2018 [Online]. Available: <https://aws.amazon.com/api-gateway/>

[900] “Amazon aurora.” Web page [Online]. Available: <https://aws.amazon.com/rds/aurora/faqs/>

[901] “Amazon cloudfront.” Web page [Online]. Available: <http://searchaws.techtarget.com/definition/Amazon-CloudFront>

[902] “Amazon cloudfront.” Web page [Online]. Available: <https://aws.amazon.com/cloudfront/>

[903] Amazon, “Introducing aws codestar.” Web page [Online]. Available: <https://aws.amazon.com/blogs/aws/new-aws-codestar/>

[904] A. AWS, “AWS deeplens.” [Online]. Available: <https://aws.amazon.com/deeplens/>

[905] aws, “Amazon dynamodb.” Web page [Online]. Available: <https://aws.amazon.com/dynamodb/>

[906] aws, “Amazon dynamodb.” Web page [Online]. Available: <https://aws.amazon.com/dynamodb/>

[907] “Amazon ec2.” Web page [Online]. Available: <https://aws.amazon.com/ec2/>

[908] AWS, “AWS elastic beanstalk.” Web page [Online]. Available: <https://aws.amazon.com/elasticbeanstalk/>

[909] AWS, “AWS elastic beanstalk faqs.” Web page [Online]. Available: <https://aws.amazon.com/elasticbeanstalk/faqs/>

[910] A. AWS, “AWS fargate.” [Online]. Available: <https://aws.amazon.com/fargate/>

[911] AWS, “Amazon glacier.” Web page [Online]. Available: <https://aws.amazon.com/glacier/>

[912] AWS, “Amazon glacier faqs.” Web page [Online]. Available: <https://aws.amazon.com/glacier/faqs/>

[913] A. AWS, “AWS lightsail.” [Online]. Available: <https://aws.amazon.com/lightsail/>

[914] “Machine learning on aws.” Web page [Online]. Available: <https://aws.amazon.com/machine-learning/>

[915] “Amazon machine learning.” Web page [Online]. Available: <https://aws.amazon.com/aml/faqs/>

[916] AWS, “Amazon rds.” Web page [Online]. Available: <https://aws.amazon.com/rds/>

[917] AWS, “Amazon rds faqs.” Web page [Online]. Available: <https://aws.amazon.com/rds/faqs/>

[918] AWS, “Amazon redshift fast and efficient data warehousing.” Web page [Online]. Available: <https://aws.amazon.com/redshift/>

[919] AWS, “Amazon redshift.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Amazon_Redshift

[920] AWS, “Amazon s3 faqs.” Web page [Online]. Available: <https://aws.amazon.com/s3/faqs/>

[921] AWS, “Amazon s3.” Web page [Online]. Available: <https://aws.amazon.com/s3/>

[922] “Amazon vpc.” Web page [Online]. Available: <https://cloudacademy.com/blog/how-and-why-to-use-vpc-for-your-amazon-aws-infrastructure/>

[923] “Ansible.” Web page [Online]. Available: http://docs.ansible.com/ansible/latest/dev_guide/overview_architecture.html

[924] Apache Software Foundation, “Apache accumulo is a sorted, distributed key/value store that provides robust, scalable data storage and retrieval.” 2018 [Online]. Available: <https://accumulo.apache.org/>

[925] “Apache ambari.” [Online]. Available: https://www.wikiwand.com/en/Apache_Ambari

[926] “Apache ambari.” [Online]. Available: <https://ambari.apache.org/>

[927] “Apache atlas.” Web page [Online]. Available: <http://atlas.apache.org/>

[928] “Apache atlas architecture.” Web page [Online]. Available: <https://atlas.apache.org/Architecture.html>

[929] “Apache avro.” Web page [Online]. Available: <http://avro.apache.org/docs/1.8.2/>

[930] “Apache chukwa.” [Online]. Available: <http://chukwa.apache.org>

[931] “Apache chukwa.” [Online]. Available: <http://chukwa.apache.org/docs/r0.5.0/>

[932] “Apache cloudstack.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_CloudStack

[933] “Apache scalability.” Web page [Online]. Available: http://tutoriaisgnulinux.com/2015/06/08/_cloudstack-4-5-centos-6-5-conceitos-basicos-e-ambiente-de-estudo-1-1/

[934] A. C. DB, “Data where you need it.” Web page [Online]. Available: <http://couchdb.apache.org>

[935] Apache, “Apache curator.” Web page [Online]. Available: <http://curator.apache.org/>

- [936] Apache, “Apache curator.” Web page [Online]. Available: <http://curator.apache.org/curator-framework/index.html>
- [937] “Apache curator.” Web page [Online]. Available: <http://nirmataoss.github.io/workflow/>
- [938] “Apache deltacloud.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/Deltacloud>
- [939] Blogs.Apache, “The apache software foundation announces apache drill as a top-level project.” Dec-2014 [Online]. Available: https://blogs.apache.org/foundation/entry/the_apache_software_foun
- [940] A. Drill, “Query any non-relational datastore (well, almost...).” 2015 [Online]. Available: <https://drill.apache.org/>
- [941] A. Drill, “Apache drill enables self-service data exploration on all data with a schema-free sql query engine.” 2017 [Online]. Available: <https://mapr.com/products/apache-drill/>
- [942] Apache, “Apache geode.” Web page [Online]. Available: <https://cwiki.apache.org/confluence/display/GEODE/Index>
- [943] Spring, “Apache geode integration with spring gemfire.” Web page [Online]. Available: <http://projects.spring.io/spring-data-gemfire/>
- [944] Spring, “Apache geode integration with spring cache.” Web page [Online]. Available: <https://docs.spring.io/spring/docs/current/spring-framework-reference/integration.html#cache>
- [945] “Apache geode integration with python.” Web page [Online]. Available: <https://github.com/gemfire/py-gemfire-rest/>
- [946] Blogs.Apache, “Database and caching platform.” Web page, 2017 [Online]. Available: <https://ignite.apache.org/>
- [947] G. Whitepapers, “Introducing apache ignite: A gridgain systems in-memory computing white paper.” Web page, 2018 [Online]. Available: <https://www.gridgain.com/resources/papers/introducing->

[apache-ignite](#)

[948] A. S. Foundation, “Apache impala.” Web page [Online]. Available: https://www.wikiwand.com/en/Apache_Impala

[949] A. S. Foundation, “Apache impala.” Web page [Online]. Available: <http://impala.apache.org/>

[950] “Apache karaf.” Web page [Online]. Available: http://karaf.apache.org/manual/latest/#_overview

[951] “Apache karaf.” Web page [Online]. Available: <https://svn.apache.org/repos/asf/karaf/site/production/manual/latest-3.0.x/jta.html>

[952] “Apache mahout.” Web page [Online]. Available: <https://mahout.apache.org/>

[953] Apache, “What is mesos? A distributed systems kernel.” 2018 [Online]. Available: <http://mesos.apache.org/>

[954] Mesos, “Why mesos?” 2018 [Online]. Available: <https://mesosphere.com/why-mesos/>

[955] Apache Software Foundation, “Apache phoenix oltp and operational analytics for apache hadoop.” 2018 [Online]. Available: <https://phoenix.apache.org/>

[956] “Apache whirr.” Web page [Online]. Available: https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/apache-whirr

[957] “Apache zookeeper.” Web page [Online]. Available: <https://www.ibm.com/analytics/hadoop/zookeeper>

[958] apatar.com, “Apatar.” Web page, 2005 [Online]. Available: <http://www.apatar.com/>

[959] “AppFog.” Web page [Online]. Available:

<https://www.ctf.io/appfog/>

[960] “AppScale.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/AppScale>

[961] “AppScale git.” Web-Github.com [Online]. Available: <https://github.com/AppScale/appscale>

[962] Apttus, “Apttus home page.” Web page [Online]. Available: <https://apttus.com/>

[963] ArangoDB, “Why arangodb.” Web page [Online]. Available: <https://www.arangodb.com/why-arangodb/>

[964] J. Steemann, “AQL: Querying a nosql database the elegant and comfortable way.” Blog, Jun-2012 [Online]. Available: <https://www.arangodb.com/2012/06/querying-a-nosql-database-the-elegant-way/>

[965] A. Walker, “Best, free and open-source database software.” Blog, Dec-2017 [Online]. Available: <https://blog.g2crowd.com/blog/database/best-free-and-open-source-database-software#arangodb>

[966] AWS, “Athena faqs.” Web page [Online]. Available: <https://aws.amazon.com/athena/faqs/>

[967] “AtomSphere.” Web page [Online]. Available: <https://boomi.com/integration/enterprise-features>

[968] “What is azure.” Web page [Online]. Available: <https://azure.microsoft.com/en-us/overview/what-is-azure/>

[969] “Azure’s choices.” Web page [Online]. Available: <https://azure.microsoft.com/en-us/overview/choose-azure-opensource/>

[970] “Azure vs aws.” Web page [Online]. Available: <https://azure.microsoft.com/en-us/overview/azure-vs-aws/>

[971] M. Wood, "An introduction to windows azure blob storage." 2013 [Online]. Available: <https://www.red-gate.com/simple-talk/cloud/cloud-data/an-introduction-to-windows-azure-blob-storage/>

[972] M. Corporation, "Microsoft azure blog: Azure cosmos db: The industry's first globally-distributed, multi-model database service." 2017 [Online]. Available: <https://azure.microsoft.com/en-us/blog/azure-cosmos-db-microsofts-globally-distributed-multi-model-database-service/>

[973] Stackify, "What is azure cosmos db? Features, benefits, pricing, and more." 2017 [Online]. Available: <https://stackify.com/what-is-azure-cosmos-db/>

[974] Backblaze, "Personal account." 2018 [Online]. Available: <https://www.backblaze.com/cloud-backup.html>

[975] Backblaze, "Business account." 2018 [Online]. Available: <https://www.backblaze.com/business-backup.html>

[976] Backblaze, "Backblaze." 2018 [Online]. Available: <https://www.backblaze.com/cloud-backup.html>

[977] A. Casalboni, "BigML offers a managed platform to build and share your datasets and models." Web page, Feb-2018 [Online]. Available: <https://cloudacademy.com/blog/bigml-machine-learning/>

[978] BigML, "About bigml." Web page, Feb-2018 [Online]. Available: <https://bigml.com/about>

[979] "Apache bigtop." Web page [Online]. Available: <http://bigtop.apache.org>

[980] "Blockchain." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Blockchain>

[981] M. Mendenhall and B. Duncan, "Bluemix is now ibm cloud: Build confidently with 170+ services." Web page, Oct-2017 [Online].

Available: <https://www.ibm.com/blogs/bluemix/2017/10/bluemix-is-now-ibm-cloud/>

[982] IBM, “IBM cloud.” Web page, 2014 [Online]. Available: <https://www.ibm.com/cloud/>

[983] BM, “IBM cloud management.” Web page, 2017 [Online]. Available: <https://www.bmc.com/it-solutions/multi-cloud-management.html>

[984] E. Shelhamer, J. Donahue, J. Long, Y. Jia, and R. Girshick, “DIY deep learning for vision: A hands-on tutorial with caffe.” 2017 [Online]. Available: https://docs.google.com/presentation/d/1UeKXVgRvxg9OUdh_UiC5G

[985] J. Yangqing et al., “Caffe: Convolutional architecture for fast feature embedding,” arXiv preprint arXiv:1408.5093, 2014.

[986] CarbonData, “CarbonData file structure.” Apache carbondata website, 2017 [Online]. Available: <https://carbondata.apache.org/file-structure-of-carbondata.html>

[987] CarbonData, “CarbonData data management.” Apache carbondata website, 2017 [Online]. Available: <https://carbondata.apache.org/data-management-on-carbondata.html>

[988] “Cascading | cascading.” Web page [Online]. Available: <https://www.cascading.org/>

[989] C. Wensel, “Cascading.” Web page [Online]. Available: <https://github.com/cwensel/cascading>

[990] “Cascading | mapr.” Web page [Online]. Available: <https://mapr.com/products/product-overview/cascading/>

[991] “CentOS project.” [Online]. Available: <https://www.centos.org/>

[992] “Clive web site.” Web page [Online]. Available:

<http://lsub.org/ls/clive.html>

[993] F. J. Ballesteros, "The clive operating system." Web page, Oct-2014 [Online]. Available: <http://lsub.org/export/clivesys.pdf>

[994] "Plan 9 from bell labs." Web page [Online]. Available: <http://plan9.bell-labs.com/plan9/about.html>

[995] "About nixos." Web page [Online]. Available: <https://nixos.org/nixos/about.html>

[996] braveclojure, "Clojure." Web page [Online]. Available: <https://www.braveclojure.com/concurrency/>

[997] Wikipedia, "Lisp." Web page [Online]. Available: [https://en.wikipedia.org/wiki/Lisp_\(programming_language\)](https://en.wikipedia.org/wiki/Lisp_(programming_language))

[998] Google, "Cloud automl-main." Web page [Online]. Available: <https://cloud.google.com/automl/>

[999] Google, "Cloud automl." Web page [Online]. Available: <https://www.blog.google/topics/google-cloud/cloud-automl-making-ai-accessible-every-business/>

[1000] "CloudHub." Web page [Online]. Available: <https://docs.mulesoft.com/runtime-manager/clouithub>

[1001] CMU, "Cloudlet vs cloud." 2018 [Online]. Available: <https://en.wikipedia.org/wiki/Cloudlet>

[1002] A. W. Services, "AWS cloudtrail logging service." 2018 [Online]. Available: <https://aws.amazon.com/cloudtrail/>

[1003] A. W. Services, "AWS cloudwatch monitoring." 2018 [Online]. Available: <https://aws.amazon.com/cloudwatch/>

[1004] Microsoft, "The microsoft cognitive toolkit." Web page [Online]. Available: <https://docs.microsoft.com/en-us/cognitive-toolkit/>

- [1005] M. Mayo, "Microsoft deep learning brings innovative features - cntk shows promise." Web page, Feb-2016 [Online]. Available: <https://www.kdnuggets.com/2016/02/microsoft-deep-learning-brings-innovative-features.html>
- [1006] A. W. Services, "AWS cognito identity provider." 2018 [Online]. Available: <https://aws.amazon.com/cognito/>
- [1007] "Azure-iot," Internet of Things - Develop an Azure-Connected IoT Solution in Visual Studio with C. [Online]. Available: <https://msdn.microsoft.com/en-us/magazine/mt694088.aspx>
- [1008] Azure, "Azure/connectthedots," Github. Jun-2017 [Online]. Available: <https://github.com/Azure/connectthedots>
- [1009] "Apache couchdb." Web page [Online]. Available: <http://couchdb.apache.org/>
- [1010] Microsoft, "Microsoft azure databricks." Web page [Online]. Available: <https://databricks.com/product/azure>
- [1011] Microsoft, "What is azure databricks." Web page [Online]. Available: <https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks>
- [1012] Google, "Google cloud datalab." Web page, Feb-2018 [Online]. Available: <https://cloud.google.com/datalab/>
- [1013] F. Lardinois, "Google launches cloud datalab, an interactive tool for exploring and visualizing data." Web page, Feb-2018 [Online]. Available: <https://techcrunch.com/2015/10/13/google-launches-cloud-datalab-an-interactive-tool-for-exploring-and-visualizing-data/>
- [1014] Datameer, "Datameer hoem page." Wb page, San Francisco, CA [Online]. Available: <https://www.datameer.com/>
- [1015] R Special Interest Group on Databases (R-SIG-DB), H. Wickham, and K. Müller, DBI: R database interface. 2017 [Online]. Available: <https://CRAN.R-project.org/package=DBI>

- [1016] RStudio, “Odbc,” Databases using R. RStudio [Online]. Available: <https://db.rstudio.com/odbc>
- [1017] H. Wickham, “Introduction to dbplyr,” Introduction to dbplyr - dbplyr. RStudio [Online]. Available: <http://dbplyr.tidyverse.org/articles/dbplyr.html>
- [1018] H. Wickham and E. Ruiz, Dbplyr: A ‘dplyr’ back end for databases. 2018 [Online]. Available: <https://CRAN.R-project.org/package=dbplyr>
- [1019] “Data virtualization overview,” Denodo. Aug-2015 [Online]. Available: <https://www.denodo.com/en/data-virtualization/overview>
- [1020] Microsoft, “Distributed machine learning toolkit.” Web page, Feb-2018 [Online]. Available: <http://www.dmtk.io/>
- [1021] G. Thomas Jr., “Microsoft open sources distributed machine learning toolkit for more efficient big data research.” Web page, Feb-2018 [Online]. Available: <https://www.microsoft.com/en-us/research/blog/microsoft-open-sources-distributed-machine-learning-toolkit-for-more-efficient-big-data-research/>
- [1022] Docker, “What is docker?” San Francisco, CA [Online]. Available: <https://www.docker.com/what-docker>
- [1023] Unknown, “What is docker?” Web page [Online]. Available: <https://opensource.com/resources/what-docker>
- [1024] dokku, “Dokku.” Web page, 2013 [Online]. Available: <http://dokku.viewdocs.io/dokku/>
- [1025] dokku, “Dokku.” Web page, 2016 [Online]. Available: <https://dokku.github.io/first-prost/welcome-to-dokku>
- [1026] W. M. Landau, Drake: Data frames in r for make. 2018 [Online]. Available: <https://CRAN.R-project.org/package=drake>
- [1027] Google, “Dremel: Interactive analysis of webscale datasets.”

Web page [Online]. Available:
<https://research.google.com/pubs/pub36632.html>

[1028] “Druid (open-source data store).” Web page [Online]. Available:
[https://en.wikipedia.org/wiki/Druid_\(open-source_data_store\)](https://en.wikipedia.org/wiki/Druid_(open-source_data_store))

[1029] Druid, “About druid.” Web page [Online]. Available:
<http://druid.io/druid.html>

[1030] F. Troßbach, “Real time fast data analytics with druid.” Blog, Aug-2016 [Online]. Available:
<https://blog.codecentric.de/en/2016/08 realtime-fast-data-analytics-druid/>

[1031] “IBM data science experience.” Web page [Online]. Available:
<https://www.ibm.com/cloud/data-science-experience>

[1032] “Optimizing cloud computing on mobile.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Edge_computing

[1033] Apache, “Apache edgent.” Apache edgent website, 2017 [Online]. Available: <http://edgent.apache.org/>

[1034] A. Edgent, “Apache edgent documentation.” Apache Edgent Website, 2017 [Online]. Available: <http://edgent.apache.org/docs/home.html>

[1035] “Elasticsearch - wikipedia.” Web page [Online]. Available:
<https://en.wikipedia.org/wiki/Elasticsearch>

[1036] “Elasticsearch: RESTful, distributed search and analytics | elastic.” Web page [Online]. Available:
<https://www.elastic.co/products/elasticsearch>

[1037] “DB-engines ranking - popularity ranking of datamase management systems.” Web page [Online]. Available: <https://db-engines.com/en/ranking>

[1038] ELK, “ELK stack.” Web Home Page, 2015 [Online]. Available:

<https://www.elastic.co/elk-stack>

[1039] AWS, “Amazon emr.” Web page [Online]. Available: <https://aws.amazon.com/emr/>

[1040] ESRI, “ESRI access and data services.” Web page, 2018 [Online]. Available: <https://developers.arcgis.com/content-and-services/>

[1041] “Ethereum.” Web page [Online]. Available: <https://www.ethereum.org/>

[1042] Google, “Firebase expands to become unified app platform.” Web page [Online]. Available: <https://firebase.googleblog.com/2016/05/firebase-expands-to-become-unified-app-platform.html>

[1043] Google, “Firebase expands to become unified app platform.” Web page [Online]. Available: <https://firebase.googleblog.com>

[1044] Google, “GCM and fcm frequently asked questions.” Web page [Online]. Available: <https://developers.google.com/cloud-messaging/faq>

[1045] Google, “Real-time collaboration with no server code.” Web page [Online]. Available: <https://firepad.io/>

[1046] I. Firepad, “Open source collaborative online editor.” Web page [Online]. Available: <http://texteditors.org/cgi-bin/wiki.pl?Firepad>

[1047] S. Vasani, “Fission: Serverless functions as a service for kubernetes.” Web page [Online]. Available: <http://blog.kubernetes.io/2017/01/fission-serverless-functions-as-service-for-kubernetes.html>

[1048] Fluentd Project, “Build your unified logging layer.” 2018 [Online]. Available: <https://www.fluentd.org/>

[1049] A. W. Services, “AWS foundation benchmarks.” 2018 [Online]. Available: <https://aws.amazon.com/blogs/security/announcing->

[industry-best-practices-for-securing-aws-resources/](#)

[1050] “Future grid.” Web page [Online]. Available: <http://www.future-grid.com.au/>

[1051] “Big data solutions - google cloud platform.” [Online]. Available: <https://cloud.google.com/solutions/big-data/>

[1052] “Cloud dataproc - cloud-native hadoop and spark - google cloud platform.” [Online]. Available: <https://cloud.google.com/dataproc/>

[1053] Google, “Google genomics.” Web page [Online]. Available: <https://cloud.google.com/genomics/>

[1054] gephi.org, “Gephi.” Web page, 2008 [Online]. Available: <https://gephi.org>

[1055] gephi.org, “Gephi.” Web page, 2008 [Online]. Available: <https://gephi.org/features/>

[1056] OpenRefine, “OpenRefine.” Web page, 2018 [Online]. Available: <https://github.com/OpenRefine/OpenRefine/wiki/Data-Sources>

[1057] GitHub.com, “GitHub developer | github developer guide.” Web page, 2018 [Online]. Available: <https://developer.github.com/>

[1058] A. Gobblin, “Apache gobblin.” Apache Gobblin Website, 2017 [Online]. Available: <http://gobblin.incubator.apache.org/>

[1059] A. Tiwari, “Gobblin as a service.” Apache Gobblin Website, Jul-2017 [Online]. Available: <https://cwiki.apache.org/confluence/display/GOBBLIN/Gobblin+as+a+service>

[1060] J. Lowry, “Designing for scale.” Web page, 2017 [Online]. Available: <https://cloud.google.com/appengine/articles/scalability>

[1061] Google, “Google cloud platform.” Web page, 2018 [Online]. Available: <https://cloud.google.com/appengine/docs/the-appengine>

environments

- [1062] Google, “Google cloud platform.” Web page, 2018 [Online]. Available: <https://cloud.google.com/appengine/docs/the-appengine-environment>
- [1063] T. Network, “Google bigquery.” [Online]. Available: <http://searchdatamanagement.techtarget.com/definition/Google-BigQuery>
- [1064] Google, “Google bigquery.” [Online]. Available: <http://bit.ly/2rGKRe2>
- [1065] Google, “CLOUD bigtable.” Web page [Online]. Available: <https://cloud.google.com/bigtable/>
- [1066] Google, “Compute engine.” Web page, 2018 [Online]. Available: <https://cloud.google.com/compute>
- [1067] Google, “Google.” 2018 [Online]. Available: <https://www.google.com/docs/about/>
- [1068] Google, “Google load balancing.” Web page [Online]. Available: <https://cloud.google.com/load-balancing/>
- [1069] Google, “Google stackdriver.” Web page [Online]. Available: <https://cloud.google.com/stackdriver/>
- [1070] A. Gossip, “Apache gossip.” Apache Gossip Website, 2017 [Online]. Available: <http://gossip.incubator.apache.org/>
- [1071] W. contributors, “Gossip protocol.” 2017 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Gossip_protocol&oldid=809583854
- [1072] Wikipedia, “Web services description language.” Web page, 2018 [Online]. Available: https://en.wikipedia.org/wiki/Web_Services_Description_Language

[1073] Wikipedia, “JSON.” Web page, 2018 [Online]. Available: <https://en.wikipedia.org/wiki/JSON>

[1074] Wikipedia, “Representational state transfer.” Web page, 2018 [Online]. Available: https://en.wikipedia.org/wiki/Representational_state_transfer

[1075] F. O. Source, “Introduction to graphql.” 2018 [Online]. Available: <http://graphql.org/>

[1076] AWS, “Amazon greengrass.” Web page [Online]. Available: <https://aws.amazon.com/greengrass/>

[1077] H2O, “H2O.” Web page [Online]. Available: <https://www.h2o.ai/h2o/>

[1078] “Apache hadoop.” Web page [Online]. Available: <http://hadoop.apache.org/>

[1079] “Apache hbase.” Web page [Online]. Available: <https://hbase.apache.org/>

[1080] L. Levenenz and A. Sears, “HCatalog.” Web (Confluence), Jan-2016 [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/HCatalog>

[1081] “Home page - hpcc systems.” [Online]. Available: <https://hpccsystems.com/>

[1082] “EHue, the self service open source analytics workbench for browsing, querying and visualizing data interactively.” Web page [Online]. Available: <http://gethue.com/sql-editor/>

[1083] “Hue (hadoop) - wikipedia.” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Hue_\(Hadoop\)](https://en.wikipedia.org/wiki/Hue_(Hadoop))

[1084] “What’s the difference between the 5 hyperledger blockchain projects?” Web page [Online]. Available: <https://www.sdxcentral.com/articles/news/whats-the-difference->

[between-the-5-hyperledger-blockchain-projects/2017/09/](#)

[1085] “Hyperledger burrow.” Web page [Online]. Available: <https://www.hyperledger.org/projects/hyperledger-burrow>

[1086] “Hyperledger fabric.” Web page [Online]. Available: <https://hyperledger.org/projects/fabric>

[1087] “Fabric 1.0: Hyperledger releases first production-ready blockchain software.” Web page [Online]. Available: <https://www.coindesk.com/fabric-1-0-hyperledger-releases-first-production-ready-blockchain-software/>

[1088] “Guest post: Phillip j. Windley, ph.d., chair, sovrin foundation.” Web page [Online]. Available: <https://www.hyperledger.org/blog/2017/05/02/hyperledger-welcomes-project-indy>

[1089] “Hyperledger iroha graduates to active status.” Web page [Online]. Available: <https://www.hyperledger.org/blog/2017/05/22/hyperledger-iroha-graduates-to-active-status>

[1090] “Hyperledger sawtooth.” Web page [Online]. Available: <https://hyperledger.org/projects/sawtooth>

[1091] “Hyperledger releases hyperledger sawtooth 1.0.” Web page [Online]. Available: <https://www.linuxfoundation.org/press-release/hyperledger-releases-hyperledger-sawtooth-1-0>

[1092] IBM, “IBM big replicate.” Web page [Online]. Available: <https://www.ibm.com/us-en/marketplace/big-replicate>

[1093] M. Mendenhall and B. Duncan, “BlueMix is now ibm cloud.” Web page, 2017 [Online]. Available: <https://www.ibm.com/blogs/bluemix/2017/10/bluemix-is-now-ibm-cloud/>

[1094] IBM, “IBM db2 big sql.” Web page [Online]. Available:

<https://www.ibm.com/us-en/marketplace/big-sql>

[1095] “ID2020.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/ID2020>

[1096] Informatica, “About the informatica.” Web page [Online]. Available: <https://www.informatica.com/products/cloud-integration.html#fbid=CaQhhmTzHcY>

[1097] Instabug, “In-app feedback and bug reporting for mobile apps.” Web page [Online]. Available: <https://instabug.com/>

[1098] Instabug, “Description about instabug.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/Instabug>

[1099] “Path to smarter technology.” Web page [Online]. Available: <https://https://www.intelcloudfinder.com/>

[1100] F. Online, “Jaspersoft review.” Web page, Dec-2017 [Online]. Available: <https://reviews.financesonline.com/p/jaspersoft/>

[1101] Wikipedia, “Big data.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Big_data

[1102] Wikipedia, “NoSQL.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/NoSQL>

[1103] Wikipedia, “Java database connectivity.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Java_Database_Connectivity

[1104] Wikipedia, “XML.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/XML>

[1105] Wikipedia, “JSON.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/JSON>

[1106] Wikipedia, “Comma-separated values.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Comma-separated_values

[1107] Wikipedia, “Hibernate.” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Hibernate_\(framework\)](https://en.wikipedia.org/wiki/Hibernate_(framework))

[1108] Wikipedia, “Plain old java object.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Plain_old_Java_object

[1109] T. Jaspersoft, “Reporting software.” Web page [Online]. Available: <https://www.jaspersoft.com/reporting-software>

[1110] Wikipedia, “JasperReports.” Web page, Dec-2017 [Online]. Available: <https://en.wikipedia.org/wiki/JasperReports>

[1111] “Wikipedia javascript.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/JavaScript>

[1112] Jelastic, “Jelastic.” Web page [Online]. Available: <https://jelastic.com/>

[1113] “JMP 9: A really new version.” Web page [Online]. Available: <https://www.scientificcomputing.com/article/2011/05/jmp-9-really-new-version>

[1114] “Statistical software | jmp software from sas.” Web page [Online]. Available: <https://wwwjmp.com/>

[1115] “JMP pro.” Web page [Online]. Available: <https://www.sas.com/jmpstore/products-solutions/jmp-pro/prodJMPPRO.html>

[1116] “Apache kafka.” Web page [Online]. Available: <https://kafka.apache.org/>

[1117] “Keras.” Web page [Online]. Available: <https://keras.io/>

[1118] PredictiveAnalyticsToday, “KNIME analytics platform.” Web page, Feb-2018 [Online]. Available: <https://www.predictiveanalyticstoday.com/knime/>

[1119] KNIME, “KNIME analytics platform.” Web page, Feb-2018

[Online]. Available: <https://www.knime.com/knime-analytics-platform>

[1120] Kubernetes, “Kubernetes.” Web page [Online]. Available: <https://kubernetes.io/>

[1121] G. A. da Costa, “KUBERNETES! Let’s take it easy...” Web page, Feb-2017 [Online]. Available: <https://medium.com/quick-mobile/kubernetes-lets-take-it-easy-924467dc5b21>

[1122] Wikipedia, “Kubernetes.” Web page, Feb-2018 [Online]. Available: <https://en.wikipedia.org/wiki/Kubernetes>

[1123] “Apache kudu - overview.” Web page [Online]. Available: <http://kudu.apache.org/overview.html>

[1124] “Apache kylin.” Web page [Online]. Available: <https://www.slideshare.net/YangLi43/apache-kylin-deep-dive-2014-dec>

[1125] “Keras.” Web page [Online]. Available: <https://github.com/Microsoft/LightGBM>

[1126] DRIVEN, “Introducing lingual-open source ansi sql for hadoop.” Web page, Feb-2013 [Online]. Available: <http://www.driven.io/2013/02/introducing-lingual-open-source-ansi-sql-for-hadoop/>

[1127] Wikipedia, “Apache maven.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_Maven

[1128] CASCADING, “Lingual.” Web page [Online]. Available: <http://www.cascading.org/projects/lingual/>

[1129] Wikipedia, “Cascading (software).” Web page [Online]. Available: [https://en.wikipedia.org/wiki/Cascading_\(software\)](https://en.wikipedia.org/wiki/Cascading_(software))

[1130] Wikipedia, “SQL.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/SQL>

- [1131] Wikipedia, “Apache hadoop.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop
- [1132] E. Sun, “Open sourcing wherehows: A data discovery and lineage portal.” Blog, Mar-2016 [Online]. Available: <https://engineering.linkedin.com/blog/2016/03/open-sourcing-wherehows--a-data-discovery-and-lineage-portal>
- [1133] “WhereHows github home.” Code Repository [Online]. Available: <https://github.com/linkedin/WhereHows/wiki>
- [1134] Linode, “About the linode.” Web page [Online]. Available: <https://www.linode.com/>
- [1135] diabarcelona, “Logicalglue.” Web page [Online]. Available: <http://www.diabarcelona.com/logical-glue-online-innovative-predictive-analytics-for-financial-services/>
- [1136] “Open source big data analytics and visualization: Lumify.” Web page [Online]. Available: <https://n0where.net/open-source-big-data-analytics-and-visualization-lumify>
- [1137] S. Gillard, “5 open source big data analytics tools to bolster your business intelligence.” Blog, Jul-2016 [Online]. Available: <https://www.linkedin.com/pulse/5-open-source-big-data-analytics-tools-bolster-your-business-gillard/>
- [1138] “Altamira’s lumify now available on amazon aws marketplace.” Web page, Jun-2017 [Online]. Available: <https://www.altamiracorp.com/index.php/2017/06/05/altamiras-lumify-now-available-on-amazon-aws-marketplace/>
- [1139] “Altamira lumify now available on microsoft azure.” Web page, Aug-2017 [Online]. Available: <https://www.altamiracorp.com/index.php/2017/08/29/altamira-lumify-now-available-on-microsoft-azure/>
- [1140] A. S. Foundation, “Mahout.” Web page, Dec-2017 [Online].

Available: <https://mahout.apache.org>

[1141] A. N. et al., "Map-reduce for machine learning on multicore." Web page [Online]. Available: <https://cs.stanford.edu/people/ang/papers/nips06-mapreducemulticore.pdf>

[1142] G. Ingersoll, "Introducing apache mahout." Web page, Sep-2009 [Online]. Available: <https://www.ibm.com/developerworks/java/library/j-mahout/>

[1143] Wikipedia, "Apache mahout." Web page, Sep-2017 [Online]. Available: https://en.wikipedia.org/wiki/Apache_Mahout

[1144] MapBox, "About map box." Web page, 2018 [Online]. Available: <https://www.mapbox.com/about/>

[1145] MariaDB, "About mariadb." Web page [Online]. Available: <https://mariadb.com/about-us>

[1146] Wikipedia, "MySQL." Web page [Online]. Available: <https://en.wikipedia.org/wiki/MySQL>

[1147] Wikipedia, "MariaDB." Web page [Online]. Available: <https://en.wikipedia.org/wiki/MariaDB>

[1148] M. Foundation, "About mariadb." Web page, Jan-2018 [Online]. Available: <https://mariadb.org/about/>

[1149] Debian, "About debian." Web page [Online]. Available: <https://www.debian.org/intro/about>

[1150] Wikipedia, "Ubuntu (operating system)." Web page [Online]. Available: [https://en.wikipedia.org/wiki/Ubuntu_\(operating_system\)](https://en.wikipedia.org/wiki/Ubuntu_(operating_system))

[1151] Wikipedia, "Linux." Web page [Online]. Available: <https://en.wikipedia.org/wiki/Linux>

[1152] MariaDB, "MariaDB server." Web page [Online]. Available:

<https://mariadb.com/products/technology/server>

[1153] Wikipedia, “XtraDB.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/XtraDB>

[1154] Concepts, “DC/os 1.10 documentation.” 2018 [Online]. Available: <https://docs.mesosphere.com/1.10/overview/concepts/>

[1155] Features, “DC/os 1.10 documentation.” 2018 [Online]. Available: <https://docs.mesosphere.com/1.10/overview/features/>

[1156] Architecture, “DC/os 1.10 documentation.” 2018 [Online]. Available: <https://docs.mesosphere.com/1.10/overview/architecture/>

[1157] “Apache metron.” Web page [Online]. Available: <https://cwiki.apache.org/confluence/display/METRON/>

[1158] “Apache metron - hortonworks.” Web page [Online]. Available: https://hortonworks.com/apache/metron/#section_3

[1159] A. Milagro, “What is apache milagro.” Apache Milagro Website, 2017 [Online]. Available: <https://milagro.apache.org/>

[1160] A. Milagro, “Milagro documentation.” Apache Milagro Website, 2017 [Online]. Available: <http://docs.milagro.io/en/>

[1161] mLab, “MLab for managed database services.” Web page [Online]. Available: <https://mlab.com/>

[1162] MonetDB, “Column store features.” Web page [Online]. Available: <https://www.monetdb.org/content/column-store-features>

[1163] DB-ENGINES, “System properties comparison monetdb vs mongodb.” Web page [Online]. Available: <https://db-engines.com/en/system/MonetDB;MongoDB>

[1164] J. Bakker, “Dutch database design drives practical innovation.” Web page, Jul-2017 [Online]. Available: <http://www.computerweekly.com/news/450422330/Dutch-database->

design-drives-practical-innovation

[1165] "Introduction to mongodb - mongodb manual 3.6." Web page [Online]. Available: <https://docs.mongodb.com/manual/introduction/>

[1166] Morpheus, "Unified ops orchestration." Web page, 2017 [Online]. Available: <https://assets.morpheusdata.com/SpudMedia/1439/attachment/Morp>

[1167] "SOA governance." Web page [Online]. Available: https://en.wikipedia.org/wiki/SOA_governance

[1168] W. Contributors, "Microsoft visual studio- wikipedia the free encyclopedia." Web page, 2018 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Microsoft_Visual_Studio

[1169] "Neo4j graph database." Web page [Online]. Available: <https://neo4j.com/product/>

[1170] R. Dillet, "Amazon introduces an aws graph database service called amazon neptune." Web page, Nov-2017 [Online]. Available: <https://techcrunch.com/2017/11/29/amazon-introduces-an-aws-graph-database-service-called-amazon-neptune/>

[1171] "Graph database." Web page [Online]. Available: <https://www.techopedia.com/definition/30577/graph-database>

[1172] J. Wong, C. Colburn, E. Meeks, and S. Vedaraman, "RAD - outlier detection on big data." Web page, Feb-2015 [Online]. Available: <https://medium.com/netflix-techblog/rad-outlier-detection-on-big-data-d6b0494371cc>

[1173] Netflix, "Netflix open source software center." Web page, 2018 [Online]. Available: <https://netflix.github.io>

[1174] "Apache nifi." Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_NiFi

[1175] "Apache nifi overview." Web page [Online]. Available:

<https://nifi.apache.org/docs/nifi-docs/html/overview.html>

[1176] “Flow-based programming.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Flow-based_programming

[1177] “Apache nifi web site.” Web page [Online]. Available: <https://nifi.apache.org/index.html>

[1178] “Node.js.” Web page [Online]. Available: <https://nodejs.org/en/about/>

[1179] J. Hester and H. Wickham, Odbc: Connect to odbc compatible databases (using the dbi interface). 2018 [Online]. Available: <https://CRAN.R-project.org/package=odbc>

[1180] Microsoft, “Microsoft’s onedrive.” 2018 [Online]. Available: <https://onedrive.live.com/about/en-us/>

[1181] A. S. Foundation, “Oozie.” Web page [Online]. Available: <http://oozie.apache.org/>

[1182] “Openchain - blockchain technology for the enterprise.” Web page [Online]. Available: <https://www.openchain.org/>

[1183] J. Redman, “Openchain: Enterprise-ready blockchain technology.” Web page [Online]. Available: <https://news.bitcoin.com/openchain-enterprise-ready-blockchain-technology/>

[1184] Y. B. Perez, “Coinprism launches open source distributed ledger.” Web page [Online]. Available: <https://www.coindesk.com/coinprism-launches-open-source-distributed-ledger/>

[1185] “OpenDaylight web site.” Web page [Online]. Available: <https://www.opendaylight.org>

[1186] E. Tittel, “CIO magazine: Understanding how sdn and nfv can work together.” Web page, Jan-2014 [Online]. Available:

<https://www.cio.com/article/2379216/business-analytics/understanding-how-sdn-and-nfv-can-work-together.html>

[1187] “SDX central: What is the.opendaylight project (odl)?” Web page [Online]. Available: <https://www.sdxcentral.com/sdn/definitions/opendaylight-project/>

[1188] “Opennebula.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/OpenNebula>

[1189] “Opennebula deployment model.” Web page [Online]. Available: https://en.wikipedia.org/wiki/OpenNebula#/media/File:OpenNebula_Logo.png

[1190] Opennn, “OpenNN documentation” [Online]. Available: http://www.opennn.net/documentation/opennn_start.html

[1191] Opennn, “OpenNN neural designer” [Online]. Available: https://en.wikipedia.org/wiki/Neural_Designer

[1192] Opennn, “Opennn” [Online]. Available: http://www.opennn.net/documentation/the_neural_network_class.htm

[1193] “OpenRefine usability.” Web page [Online]. Available: <https://www.computerworld.com/article/2507728/enterprise-applications/enterprise-applications-22-free-tools-for-data-visualization-and-analysis.html?page=2>

[1194] “OpenVZ introduction.” Web page [Online]. Available: <http://www.linuxzhaji.com/?p=453>

[1195] Oracle, “Delivering hadoop, spark and data science with oracle security and cloud simplicity.” [Online]. Available: https://cloud.oracle.com/en_US/big-data

[1196] Oracle, “Oracle coherence.” Web page [Online]. Available: https://docs.oracle.com/cd/E24290_01/coh.371/e22840/definingdatag

[1197] “Oracle coherence.” Web page [Online]. Available:

https://en.wikipedia.org/wiki/Oracle_Coherence

[1198] Oracle, “Oracle coherence integration with spring framework.” Web page [Online]. Available: https://docs.oracle.com/cd/E18686_01/coh.37/e18691/integratespring

[1199] “Oracle nosql database.” Web page [Online]. Available: <http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/index.html>

[1200] Orange, “Data mining fruitful and fun.” Web page, Feb-2018 [Online]. Available: <https://orange.biolab.si/>

[1201] Wikipedia, “Orange (software).” Web page, Feb-2018 [Online]. Available: https://en.wikipedia.org/wiki/Orange_%28software%29

[1202] OrientDB, “Why-orientdb” [Online]. Available: <https://orientdb.com/why-orientdb/>

[1203] OrientDB, “OrientDB in the cloud” [Online]. Available: <https://orientdb.com/docs/last/SB-Tree-index.html>

[1204] OrientDB, Available: https://www.tutorialspoint.com/orientdb/orientdb_basic_concepts.htm

[1205] OrientDB, “OrientDB in the cloud” [Online]. Available: <https://orientdb.com/graph-database/>

[1206] ownCloud, “Owncloud for efficient file hosting services.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/OwnCloud>

[1207] Paxata, “Paxata.” Web page [Online]. Available: <https://www.paxata.com/product/self-service-data-prep/>

[1208] kdnuggets, “Paxata.” Web page [Online]. Available: <https://www.kdnuggets.com/2014/03/paxata-automates-data-preparation-for-big-data-analytics.html>

[1209] D. Dietrich, B. Heller, and B. Yang, “Data science & big data

analytics: Discovering, analyzing, visualizing and presenting data." EMC Education Services, John Wiley & Sons, Inc, 2015.

[1210] "Welcome to apache pig." Web page [Online]. Available: <https://pig.apache.org/>

[1211] "Difference between pig and hive-the two key components of hadoop ecosystem." Web page, Oct-2014 [Online]. Available: <https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79>

[1212] Pivotal, "Pivotal cloud foundry monitoring." 2017 [Online]. Available: <https://signalfx.com/pivotal-cloud-foundry-monitoring/>

[1213] B. Darrow, "Pivotal cloud foundry is not just for new apps anymore." 2016 [Online]. Available: <http://fortune.com/2016/06/06/pivotal-cloud-foundry-legacy-apps/>

[1214] vmware, "Pivotal rabbitmq vmware products." Web page, 2018 [Online]. Available: <https://www.vmware.com/products/pivotal-rabbitmq.html>

[1215] B. Borges, Pool: Object pooling. 2017 [Online]. Available: <https://CRAN.R-project.org/package=pool>

[1216] A. S. Foundation, "Apache predictionio." Web page [Online]. Available: <https://predictionio.apache.org/#what-is-apache-predictionio>

[1217] Presto, "Presto distributed sql engine for big data." Web page [Online]. Available: <https://prestodb.io>

[1218] PubNub, "PubNub data stream network." Web page [Online]. Available: <https://www.pubnub.com/>

[1219] Wikipedia, "Description about pubnub." Web page [Online]. Available: <https://en.wikipedia.org/wiki/PubNub>

[1220] "Apache pulsar." Web page [Online]. Available:

<https://pulsar.incubator.apache.org/docs/latest/getting-started/ConceptsAndArchitecture/>

[1221] Puppet, “Puppet.” Web page [Online]. Available: <https://puppet.com/blog/managing-kubernetes-configuration-puppet>

[1222] PyTorch, “About torch” [Online]. Available: <http://pytorch.org/about/>

[1223] PyTorch, “Torch tutorial” [Online]. Available: <https://towardsdatascience.com/pytorch-tutorial-distilled-95ce8781a89c>

[1224] PyTorch, “Learning pytorch with examples” [Online]. Available: <https://www.analyticsvidhya.com/blog/2018/02/pytorch-tutorial/>

[1225] “QDS - big data platform - qubole.” [Online]. Available: <https://www.qubole.com/products/qubole-data-service/>

[1226] Rabbitmq, “RabbitMQ” [Online]. Available: <https://www.rabbitmq.com/>

[1227] Rabbitmq, “Clustering guide” [Online]. Available: <https://www.rabbitmq.com/clustering.html>

[1228] Rabbitmq, “Rabbitmq authentication” [Online]. Available: <https://www.rabbitmq.com/authentication.html>

[1229] Rabbitmq, “Authentication” [Online]. Available: <https://en.wikipedia.org/wiki/RabbitMQ>

[1230] Rabbitmq, “Management plugin” [Online]. Available: <http://www.rabbitmq.com/management.html>

[1231] R. Inc, “Rackspace for efficient cloud management services.” Web page [Online]. Available: <https://www.rackspace.com/en-us/about>

[1232] “Apache ranger.” Web page [Online]. Available:

<http://ranger.apache.org/>

[1233] RapidMiner, “About the rapidminer.” Web page [Online]. Available: <https://rapidminer.com/>

[1234] Redis, “Introduction to redis” [Online]. Available: <https://redis.io/topics/introduction>

[1235] Redis, “Redis modules” [Online]. Available: <https://redis.io/topics/lru-cache>

[1236] Redis, “Redis modules” [Online]. Available: <https://redis.io/topics/replication>

[1237] “RightScale cloud management.” Web page [Online]. Available: <https://www.networkworld.com/article/2164791/cloud-computing>

[1238] “Ripple.” Web page [Online]. Available: <https://imtconferences.com/ripple/>

[1239] “Global payment.” Web page [Online]. Available: https://ripple.com/files/ripple_solutions_guide.pdf

[1240] AWS, “Amazon sagemaker.” Web page [Online]. Available: <https://aws.amazon.com/sagemaker/>

[1241] “SalesCloud.” Web page [Online]. Available: <https://www.salesforce.com/products/sales-cloud/features>

[1242] “Apache samoa web site.” Web page [Online]. Available: <https://samoa.incubator.apache.org/>

[1243] A. Bifet, “Mining big data streams with apache samoa,” in Proceedings of the 6th international conference on mining ubiquitous and social environments, 2015, vol. 1521, pp. 55–55 [Online]. Available: <http://dl.acm.org/citation.cfm?id=3053868.3053878>

[1244] G. D. F. Morales and A. Bifet, “SAMOA: Scalable advanced massive online analysis,” Journal of Machine Learning Research, vol.

16, pp. 149–153, Jan. 2015 [Online]. Available: <http://jmlr.org/papers/v16/morales15a.html>

[1245] J. Chathuranga, “ML tools (java).” Blog, Nov-2016 [Online]. Available: <https://medium.com/techco/ml-tools-java-91e9b8225cf7>

[1246] F. Pedregosa et al., “Scikit-learn: Machine learning in python,” CoRR, vol. abs/1201.0490, 2012 [Online]. Available: <http://arxiv.org/abs/1201.0490>

[1247] “Facebookarchive | scribe.” Web page [Online]. Available: <https://github.com/facebookarchive/scribe>

[1248] R. Johnson, “Facebook’s technology now open source.” Web page [Online]. Available: https://www.facebook.com/note.php?note_id=32008268919

[1249] SETI, “About.” 2018 [Online]. Available: https://setiathome.berkeley.edu/sah_about.php

[1250] BOINC, “Projects.” 2018 [Online]. Available: <https://boinc.berkeley.edu/projects.php>

[1251] LHC, “History.” 2018 [Online]. Available: <http://lhcatome.web.cern.ch/about/history>

[1252] ShareLatex, “Documentation.” 2018 [Online]. Available: <https://www.sharelatex.com/learn>

[1253] ShareLatex, “Plans.” 2018 [Online]. Available: <https://www.sharelatex.com/user/subscription/plans>

[1254] Microsoft, “SharePoint.” 2018 [Online]. Available: <https://products.office.com/en-us/sharepoint/collaboration/>

[1255] S. team, “SharePoint.” 2012 [Online]. Available: <https://blogs.office.com/en-us/2012/07/17/the-new-sharepoint/>

[1256] “Skytap.” Web page [Online]. Available:

<https://www.skytap.com/>

[1257] Wikipedia, “Apache lucene.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_Lucene

[1258] tutorialspoint, “Apache solr-overview.” Web page [Online]. Available:

https://www.tutorialspoint.com/apache_solr/apache_solr_overview.htm

[1259] Wikipedia, “Apache solr.” Web page [Online]. Available: https://en.wikipedia.org/wiki/Apache_Solr

[1260] Apache, “Apache solr.” Web page [Online]. Available: <http://lucene.apache.org/solr/>

[1261] “Spagobi,” 100 open source Business Intelligence. [Online]. Available: <http://www.spagobi.org/product/>

[1262] “SpagoBI open source,” SpagoBI Open Source - Stratebi. [Online]. Available: http://www.stratebi.com/en_GB/spagobi#big-data

[1263] Google, “Google spanner.” Web page [Online]. Available: <https://cloud.google.com/spanner/>

[1264] Spinnaker, “Spinnaker io.” Web page [Online]. Available: <https://www.spinnaker.io/>

[1265] Google, “Spinnaker.” Web page [Online]. Available: <https://opensource.google.com/projects/spinnaker>

[1266] “About sqlite architecture.” Web page [Online]. Available: <https://sqlite.org/arch.html>

[1267] “Sqoop user guide (v1.4.7).” Web page [Online]. Available: http://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html#_introduction

[1268] “Stardog 5 the manual.” Web page [Online]. Available: <https://www.stardog.com/docs/>

[1269] K. Cerans, G. Barzdins, R. Liepins, J. Ovcinnikova, S. Rikacovs, and A. Sprogis, "Graphical schema editing for stardog owl/rdf databases using owlgred/s." in OWLED, 2012, vol. 849 [Online]. Available: <http://dblp.uni-trier.de/db/conf/owled/owled2012.html#CeransBLORS12>

[1270] S. Carey, "Graph database vendors: Who they are, what they do and who their customers are." Web page, Aug-2017 [Online]. Available: <https://www.computerworlduk.com/galleries/data/graph-database-vendors-who-are-they-what-do-they-do-who-their-customers-are-3639961>

[1271] The MITRE Corporation, "Synthetic patient generation." 2018 [Online]. Available: <https://synthetichealth.github.io/synthea/>

[1272] N. Patki, "The synthetic data vault: Generative modeling for relational databases," Master's thesis, Massachusetts Institute of Technology, 2016.

[1273] A. SystemML, "Apache systemml," Apache SystemML - Declarative Large-Scale Machine Learning. [Online]. Available: <http://systemml.apache.org/>

[1274] "Apache systemml blog," The Apache Software Foundation Blog. [Online]. Available: <https://blogs.apache.org/foundation/entry/the-apache-software-foundation-announces13>

[1275] Tableau, "Tableau home page." Web page, Seattle, Washington [Online]. Available: <https://www.tableau.com/>

[1276] Talend, "Talend." Web page [Online]. Available: <https://www.talend.com>

[1277] TensorFlow, "TensorFlow home page" [Online]. Available: <https://www.tensorflow.org/>

[1278] TensorFlow, "TensorFlow high level apis" [Online]. Available:

https://www.tensorflow.org/guide/#high_level_apis

[1279] TensorFlow, “TensorFlow summaries and tensorboard” [Online]. Available:

https://www.tensorflow.org/programmers_guide/summaries_and_tensorboard

[1280] “Teradata intelliflex.” Web page [Online]. Available:

<https://www.teradata.com/products-and-services/intelliflex-features>

[1281] Teradata, “Teradata intellibase.” Web page [Online]. Available:

<https://www.teradata.com/products-and-services/intellibase/>

[1282] Teradata, “Teradata kylo.” Web page [Online]. Available:

<https://kylo.io/index.html>

[1283] Wikipedia, “Theano.” Web page, Feb-2018 [Online]. Available:

[https://en.wikipedia.org/wiki/Theano_\(software\)](https://en.wikipedia.org/wiki/Theano_(software))

[1284] Theano, “Theano.” Web page [Online]. Available:

<http://deeplearning.net/software/theano/>

[1285] Google, “The go programming language.” [Online]. Available:

<https://golang.org/>

[1286] “Tibco datasynapse gridserver.” Web page [Online]. Available:

<http://enacademic.com/dic.nsf/enwiki/5982527>

[1287] “Tibco datasynapse gridserver.” Web page [Online]. Available:

<https://www.tibco.com/press-releases/2009/tibco-software-acquires-datasynapse>

[1288] “Tibco datasynapse gridserver.” Web page [Online]. Available:

<https://docs.tibco.com/products/tibco-datasynapse-gridserver-6-2-0>

[1289] Wikipedia, “TokuDB.” Web page, Nov-2017 [Online]. Available:

<https://en.wikipedia.org/wiki/TokuDB>

[1290] Percona, “TokuDB introduction.” Web page [Online]. Available:

<https://www.percona.com/doc/percona-tokudb-introduction>

[server/LATEST/tokudb/tokudb_intro.html](#)

[1291] Percona, “Percona tokudb.” Web page [Online]. Available: <https://www.percona.com/software/mysql-database/percona-tokudb>

[1292] A. Krmelj, “Converting your innodb mysql server to tokudb (how we did it).” Web page, Oct-2013 [Online]. Available: <http://blackbird.si/converting-your-innodb-mysql-server-to-tokudb-how-we-did-it/>

[1293] T. Data, “About the treasuredata.” Web page [Online]. Available: <https://www.treasuredata.com/>

[1294] Twilio, “Twilio.” Web page [Online]. Available: <https://en.wikipedia.org/wiki/Twilio>

[1295] CFPB, “About the consumer financial protection bureau.” Web page, 2018 [Online]. Available: <https://www.consumerfinance.gov>

[1296] Google, “Google vision api.” Web page [Online]. Available: <https://cloud.google.com/vision/>

[1297] machinelearningmastery, “Weka.” Web page [Online]. Available: <https://machinelearningmastery.com/what-is-the-weka-machine-learning-workbench/>

[1298] T. W. Bank, “World bank open data.” Web page, 2018 [Online]. Available: <https://data.worldbank.org/>

[1299] WSO2, “WSO2 analytics - features.” Web page [Online]. Available: <https://wso2.com/analytics/features>

[1300] WSO2, “Stream processing and complex event processing engine.” Code Repository [Online]. Available: <https://github.com/wso2/siddhi>

[1301] WSO2, “[WSO2Con usa 2017] driving insights for your digital business with analytics.” Web page, Mar-2017 [Online]. Available: <https://wso2.com/library/conference/2017/2/wso2con-usa-2017->

[deriving-insights-for-your-digital-business-with-analytics/](#)

[1302] WSO2, "Six business benefits of smart analytics." Web page, Feb-2017 [Online]. Available:

<https://wso2.com/library/articles/2017/02/six-business-benefits-of-smart-analytics/>

[1303] "XGBoost." Web page [Online]. Available:
<https://www.wikiwand.com/en/Xgboost>

[1304] J. Brownlee, "XGBoost." Web page [Online]. Available:
<https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

[1305] "XGBoost." Web page [Online]. Available:
<https://github.com/dmlc/xgboost>

[1306] packtpub, "Zepplin." Web page [Online]. Available:
https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/apache-zeppelin

[1307] Apache, "ApacheSpark." Web page [Online]. Available:
<https://spark.apache.org>

[1308] "Zmanda." Web page [Online]. Available:
<https://www.crunchbase.com/organization/zmanda>

[1309] "Amanda enterprise." Web page [Online]. Available:
<http://www.zmanda.com/amanda-enterprise-edition.html>

[1310] "Amanda enterprise and zmanda recovery manager." Web page [Online]. Available: <http://zmanda.com/webinars.html>