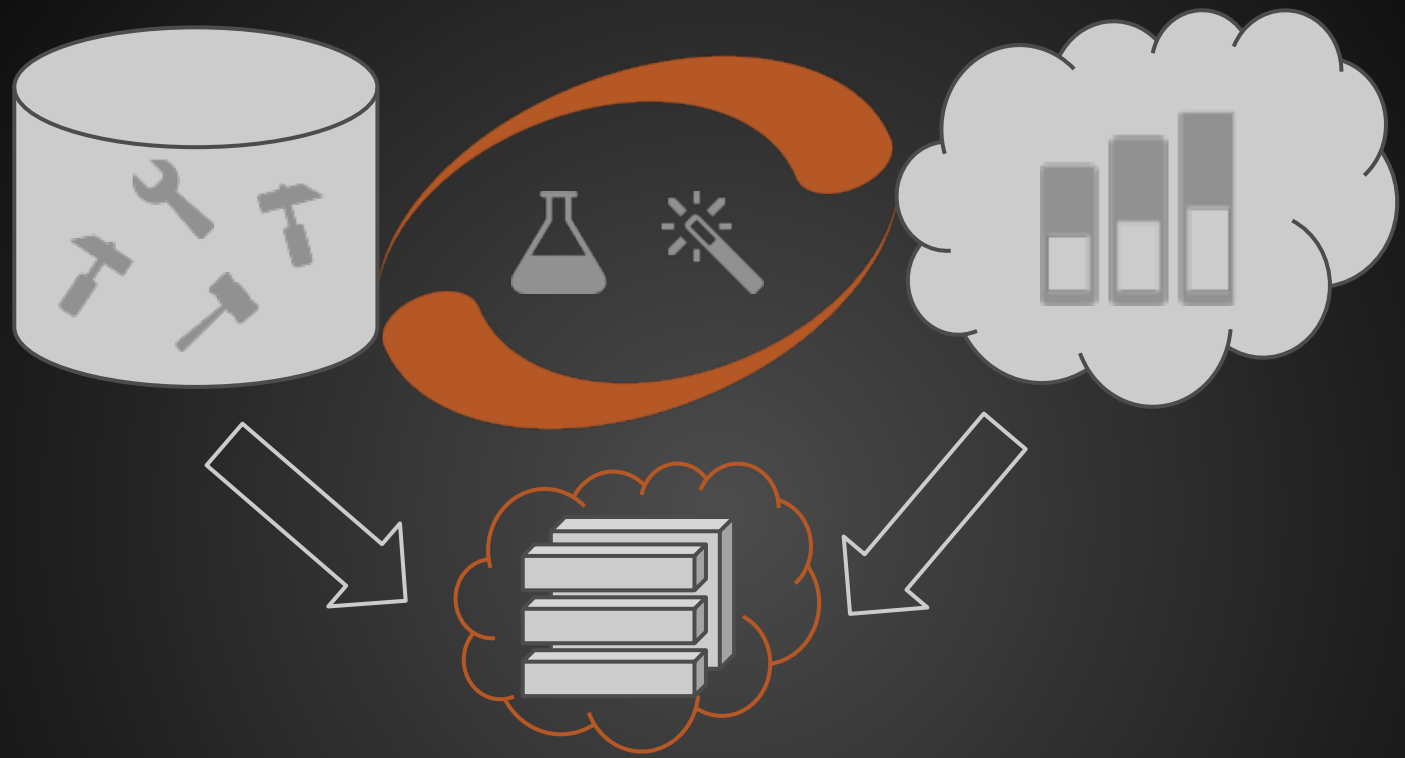
Deciphering Big Data Stacks: An Overview of Big Data Tools

Tomislav Lipic[1], Karolj Skala[1], Enis Afgan[1,2]

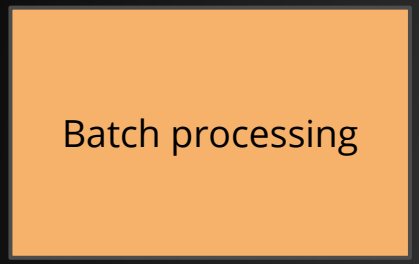
[1]Ruđer Bošković Institute (RBI)
[2]Johns Hopkins University

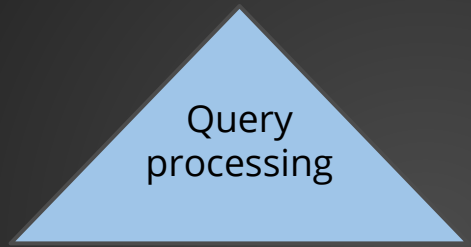
Big Data Analytics: Challenges and Opportunities (BDAC-14)
New Orleans, LA
Nov 2014

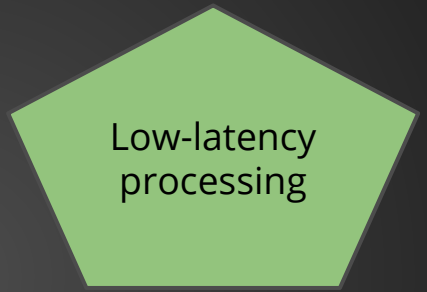


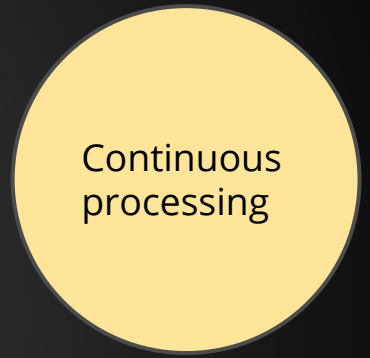
1. Tool catalog
2. Functional tool comparison
3. Tool deployment dependency graph

Existing Technologies as Trends









MapReduce
- Hadoop
- Cloudera
- Hortonworks
- MapR
- AWS EMR

HDFS
- Replicated data store

SQL-like interface
- Pig
- Hive

NoSQL databases
- Scalable storage
- Query interfaces
- Document, Key-value, Columnar, Graph

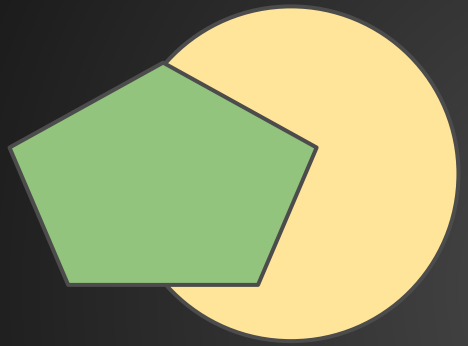
Near real-time query
- Impala
- Drill
- Dremel

Data store
- HDFS
- NoSQL

Stream processing
- Storm
- S4
- Samza
- AWS Kinesis

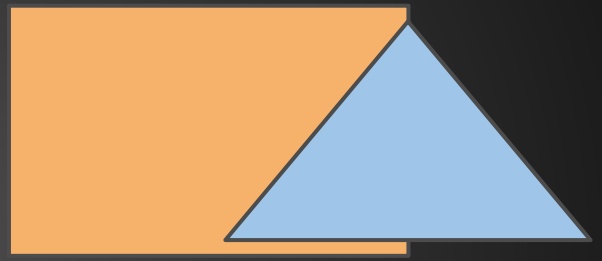
Data delivery
- Kafka
- Flume

Crossing the Trends



Spark
- Low-latency queries
- Data streams

Stratosphere Nephele
- Additional processing operators: join, union

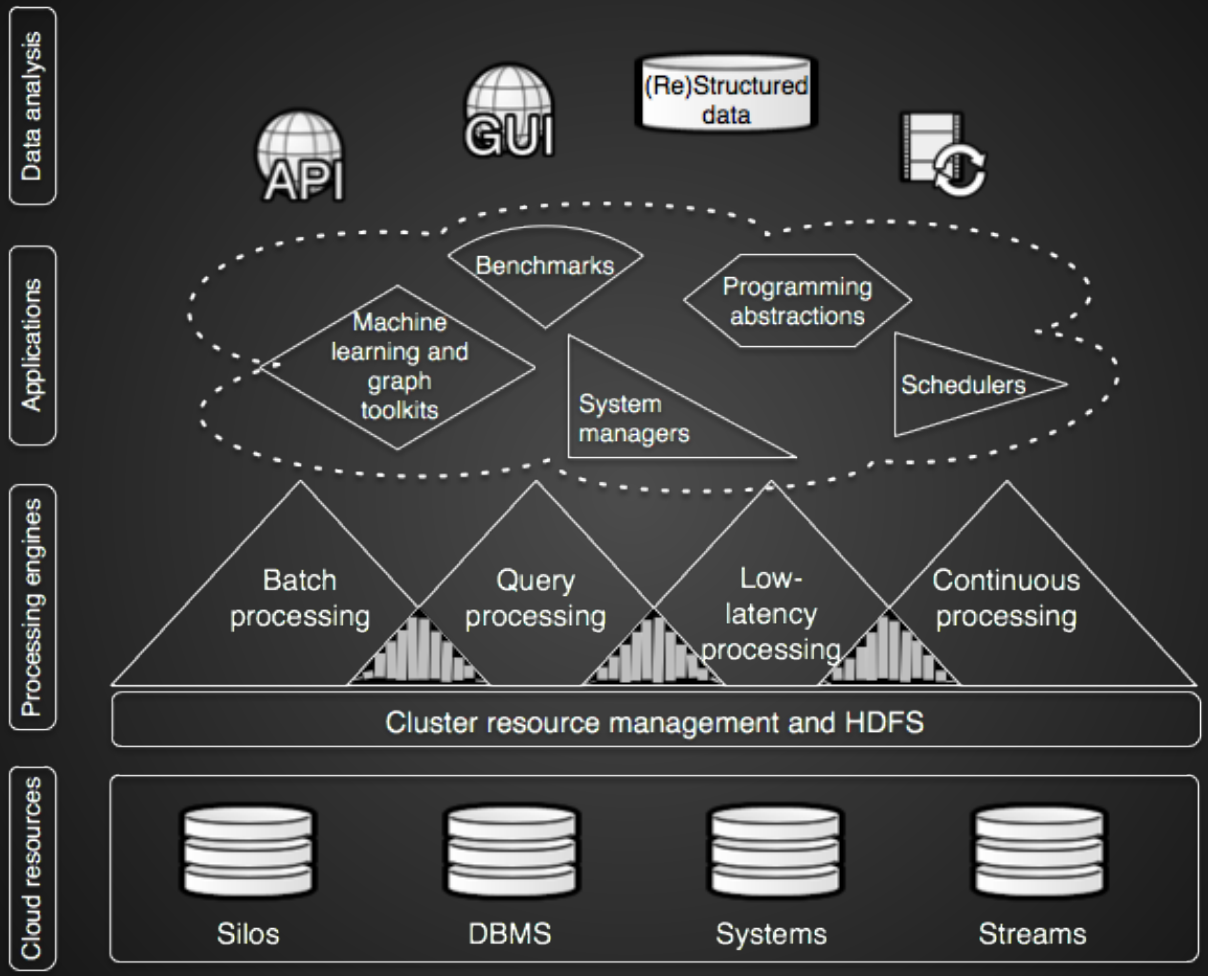


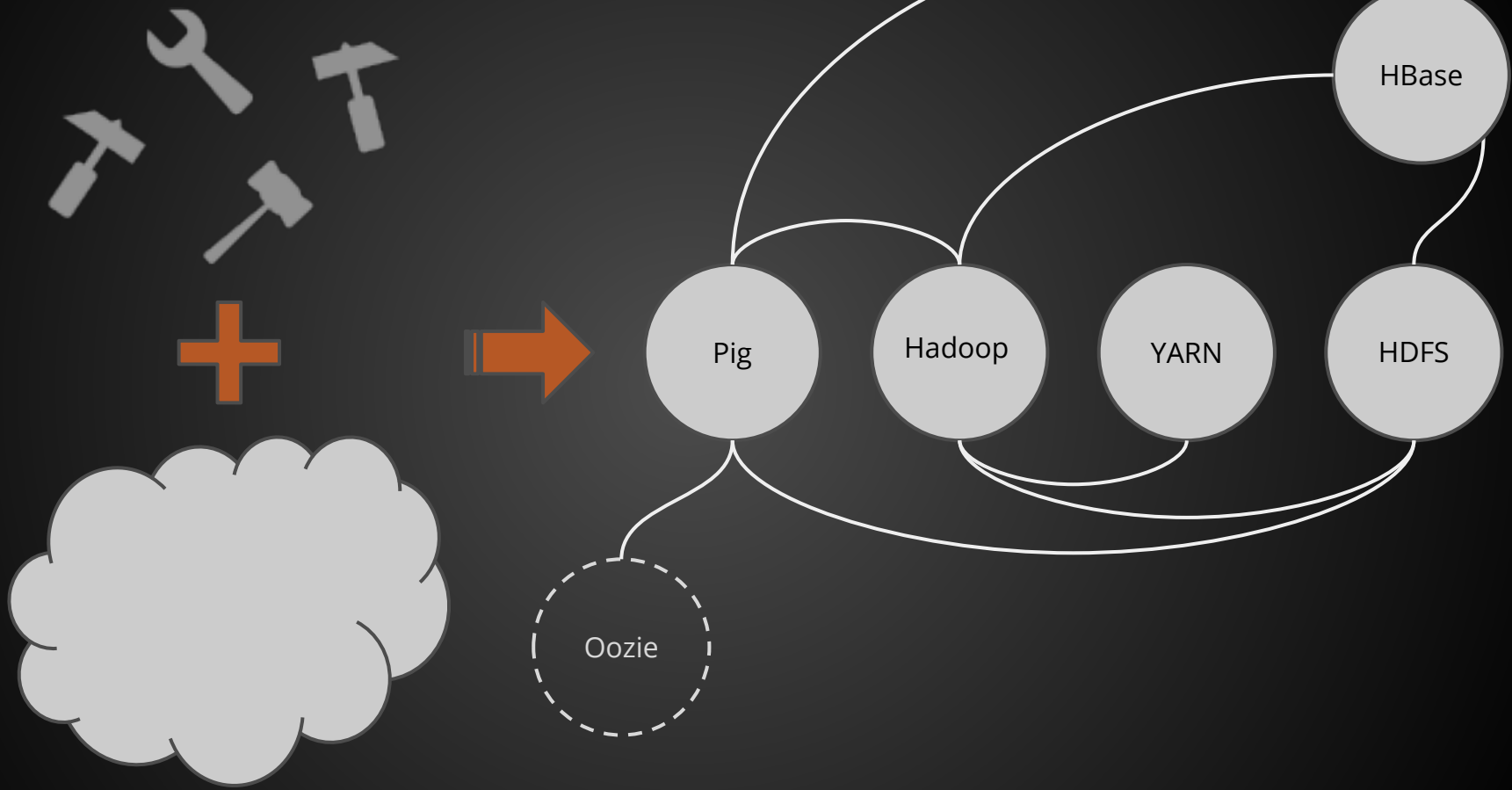
Meta-resource managers
- YARN
- Mesos

One cluster per datacenter
- On-demand resource provisioning

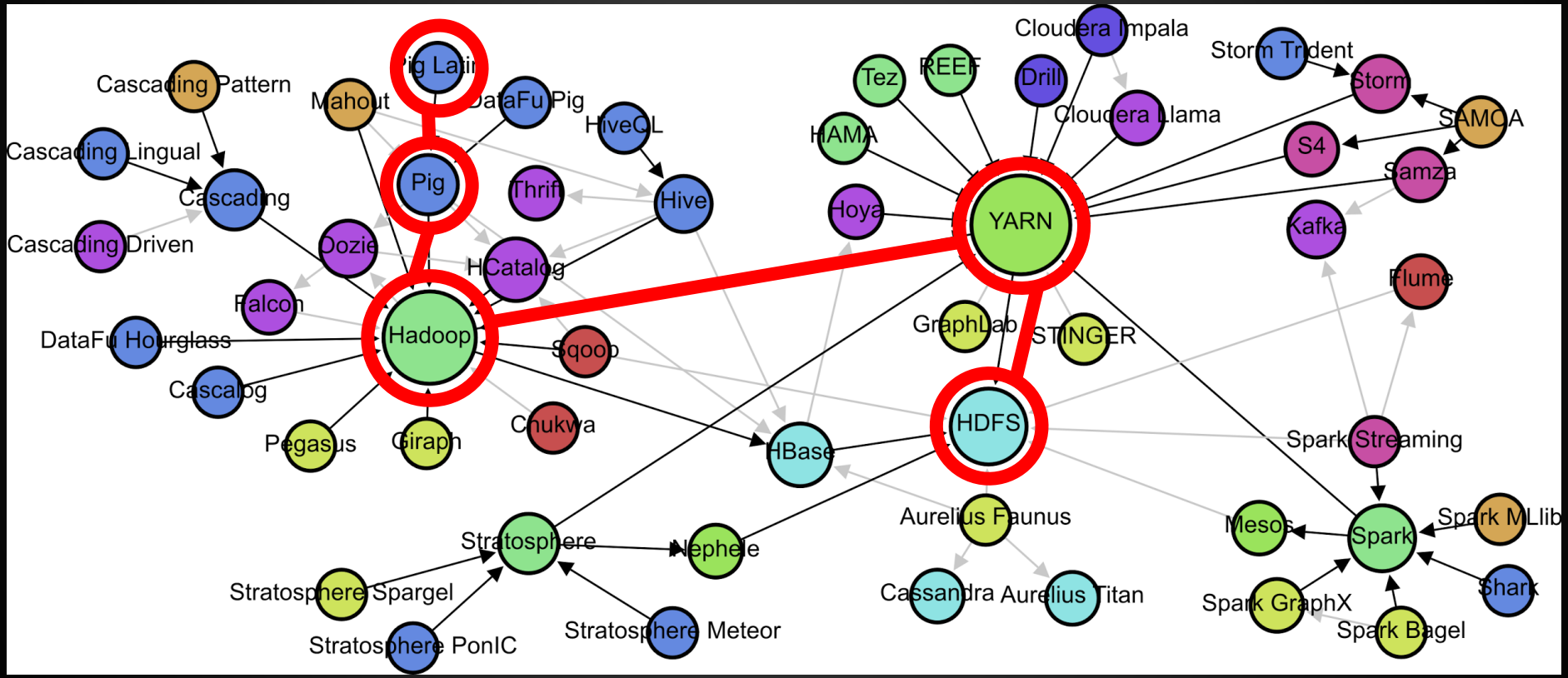
Domain-specific applications and libraries

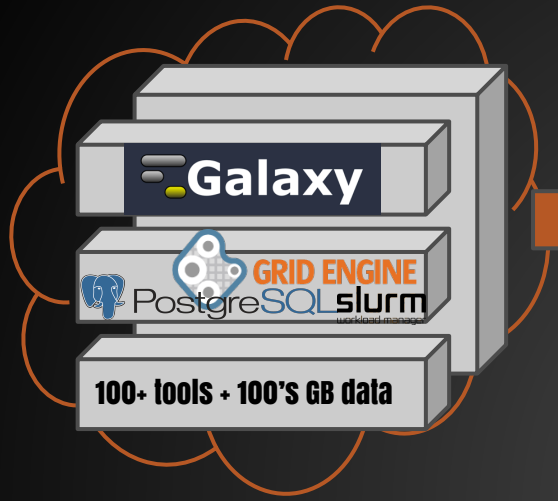
Functional classification	Existing Big Data Technologies
High-level programming abstractions	Apache Pig, Apache DataFu, Cascading Lingual, Cascalog
	Shark (for Spark), Trident (for Storm), Meteor, Sopremo and PonIC (for Stratosphere)
Big Data-aware machine-learning toolkits	Apache Mahout (for Hadoop), MLlib (for Spark), Cascading Pattern, GraphLab framework, Yahoo SAMOA
Graph processing systems	Apache Giraph (for Hadoop), Bagel (for Spark), Stratosphere Spargel, GraphX (for Spark), Pegasus, Aurelius Faunus, GraphLab PowerGraph
	Stinger, Neo4j, Aurelius Titan
Data ingestion and scheduling systems	Apache Sqoop, Apache Chukwa, Apache Flume
	Apache Falcon, Apache Oozie
Systems management solutions	Apache Hue
	Apache Ambari, Apache Helix, Apache Whirr, Cask Coopr
Benchmarking and testing applications	Berkeley Big Data benchmark, BigBench, BigDataBench,, Big Data Top 100, Apache Bigtop

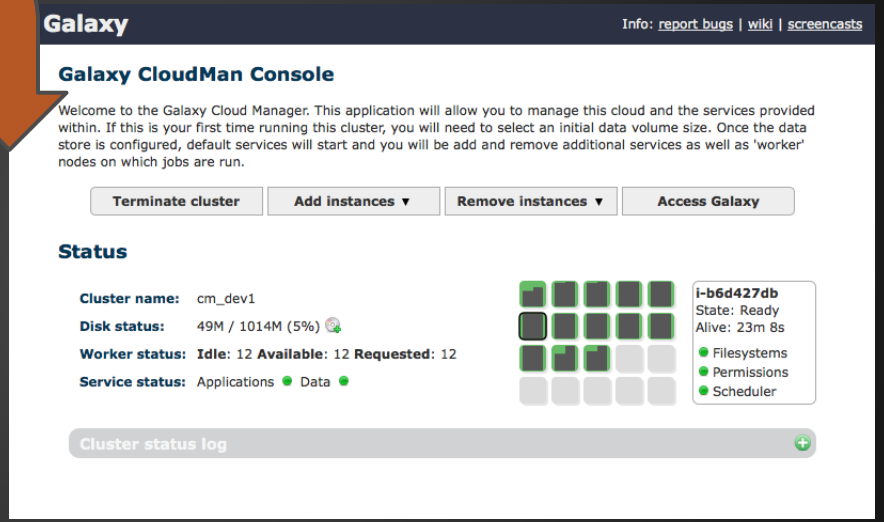

Data analysis
Applications
Processing engines
Cloud resources
API
GUI
(Re)Structured data
Benchmarks
Programming abstractions
Machine learning and graph toolkits
System managers
Schedulers
Batch processing
Query processing
Low-latency processing
Continuous processing
Cluster resource management and HDFS
Silos
DBMS
Systems
Streams


HBase
Pig
Hadoop
YARN
HDFS
Oozie

Tool Interdependency Graph




Galaxy
GRID ENGINE
PostgreSQL
slurm
workload manager
100+ tools + 100's GB data


Galaxy
Info: report bugs | wiki | screencasts
Galaxy CloudMan Console
Welcome to the Galaxy Cloud Manager. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.
Terminate cluster
Add instances ▾
Remove instances ▾
Access Galaxy
Status
Cluster name: cm_dev1
Disk status: 49M / 1014M (5%)
Worker status: Idle: 12 Available: 12 Requested: 12
Service status: Applications Data
i-b6d427db
State: Ready
Alive: 23m 8s
Filesystems
Permissions
Scheduler
Cluster status log

CloudMan

Questions?

Slides at slideshare.net/afgane
Paper at bit.ly/BDTools

Support from Ruđer Bošković Institute and Johns Hopkins University