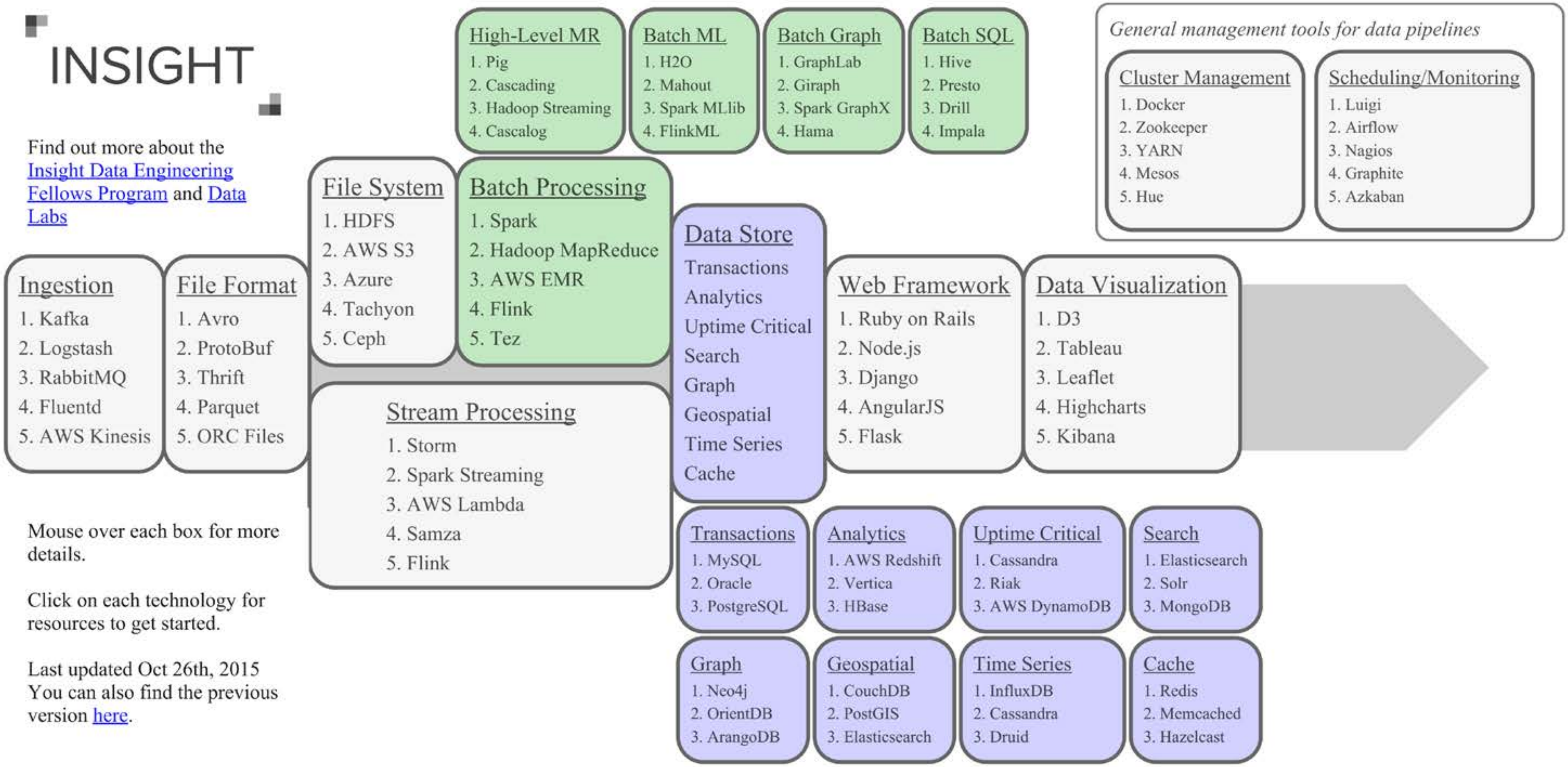
INSIGHT

Find out more about the Insight Data Engineering Fellows Program and Data Labs



Ingestion
1. Kafka
2. Logstash
3. RabbitMQ
4. Fluentd
5. AWS Kinesis

File Format
1. Avro
2. ProtoBuf
3. Thrift
4. Parquet
5. ORC Files

File System
1. HDFS
2. AWS S3
3. Azure
4. Tachyon
5. Ceph

Batch Processing
1. Spark
2. Hadoop MapReduce
3. AWS EMR
4. Flink
5. Tez

High-Level MR
1. Pig
2. Cascading
3. Hadoop Streaming
4. Cascalog

Batch ML
1. H2O
2. Mahout
3. Spark MLlib
4. FlinkML

Batch Graph
1. GraphLab
2. Giraph
3. Spark GraphX
4. Hama

Batch SQL
1. Hive
2. Presto
3. Drill
4. Impala

Stream Processing
1. Storm
2. Spark Streaming
3. AWS Lambda
4. Samza
5. Flink

Data Store
Transactions
Analytics
Uptime Critical
Search
Graph
Geospatial
Time Series
Cache

Transactions
1. MySQL
2. Oracle
3. PostgreSQL

Analytics
1. AWS Redshift
2. Vertica
3. HBase

Uptime Critical
1. Cassandra
2. Riak
3. AWS DynamoDB

Search
1. Elasticsearch
2. Solr
3. MongoDB

Graph
1. Neo4j
2. OrientDB
3. ArangoDB

Geospatial
1. CouchDB
2. PostGIS
3. Elasticsearch

Time Series
1. InfluxDB
2. Cassandra
3. Druid

Cache
1. Redis
2. Memcached
3. Hazelcast

Web Framework
1. Ruby on Rails
2. Node.js
3. Django
4. AngularJS
5. Flask

Data Visualization
1. D3
2. Tableau
3. Leaflet
4. Highcharts
5. Kibana

General management tools for data pipelines

Cluster Management
1. Docker
2. Zookeeper
3. YARN
4. Mesos
5. Hue

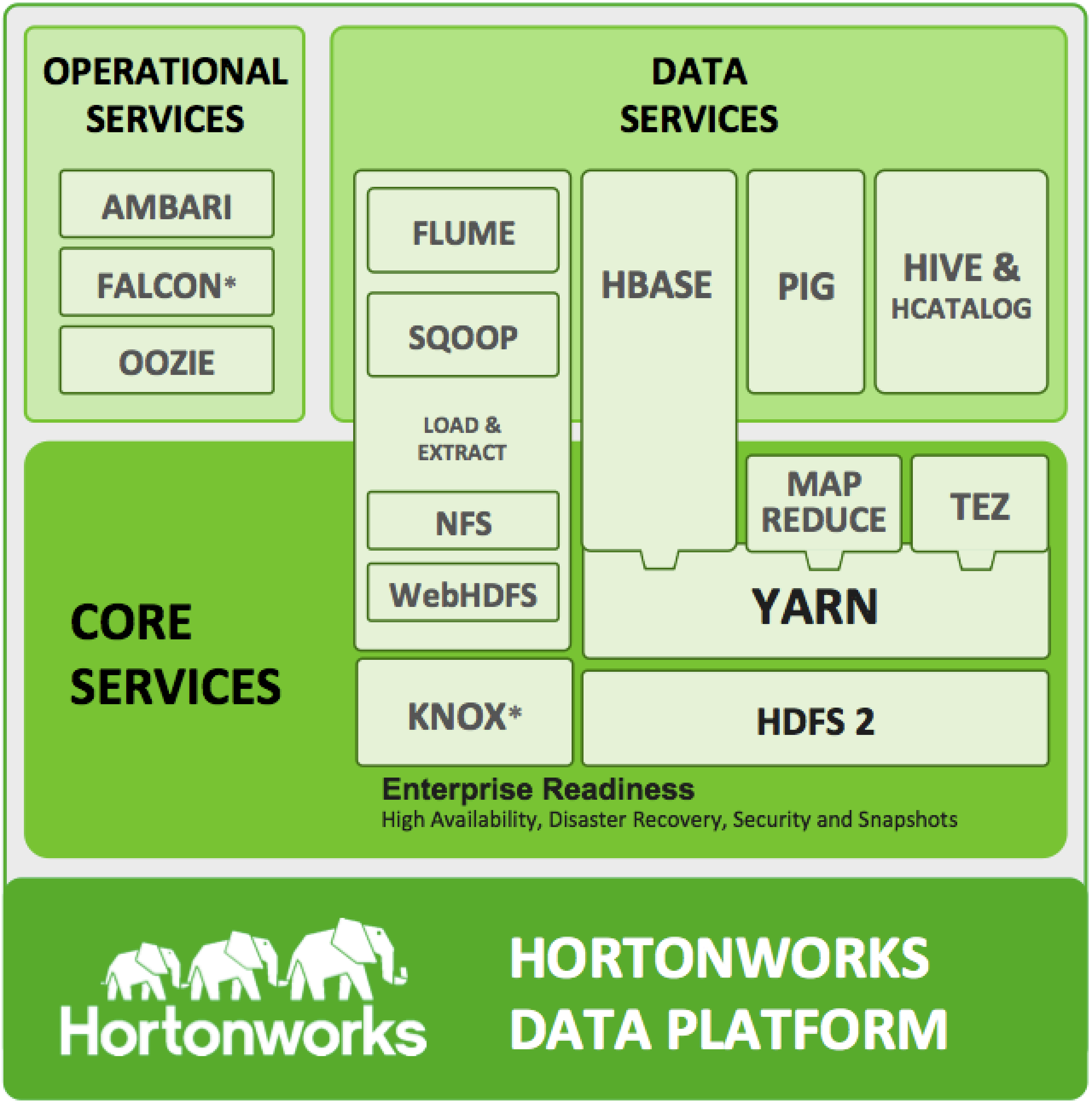
Scheduling/Monitoring
1. Luigi
2. Airflow
3. Nagios
4. Graphite
5. Azkaban

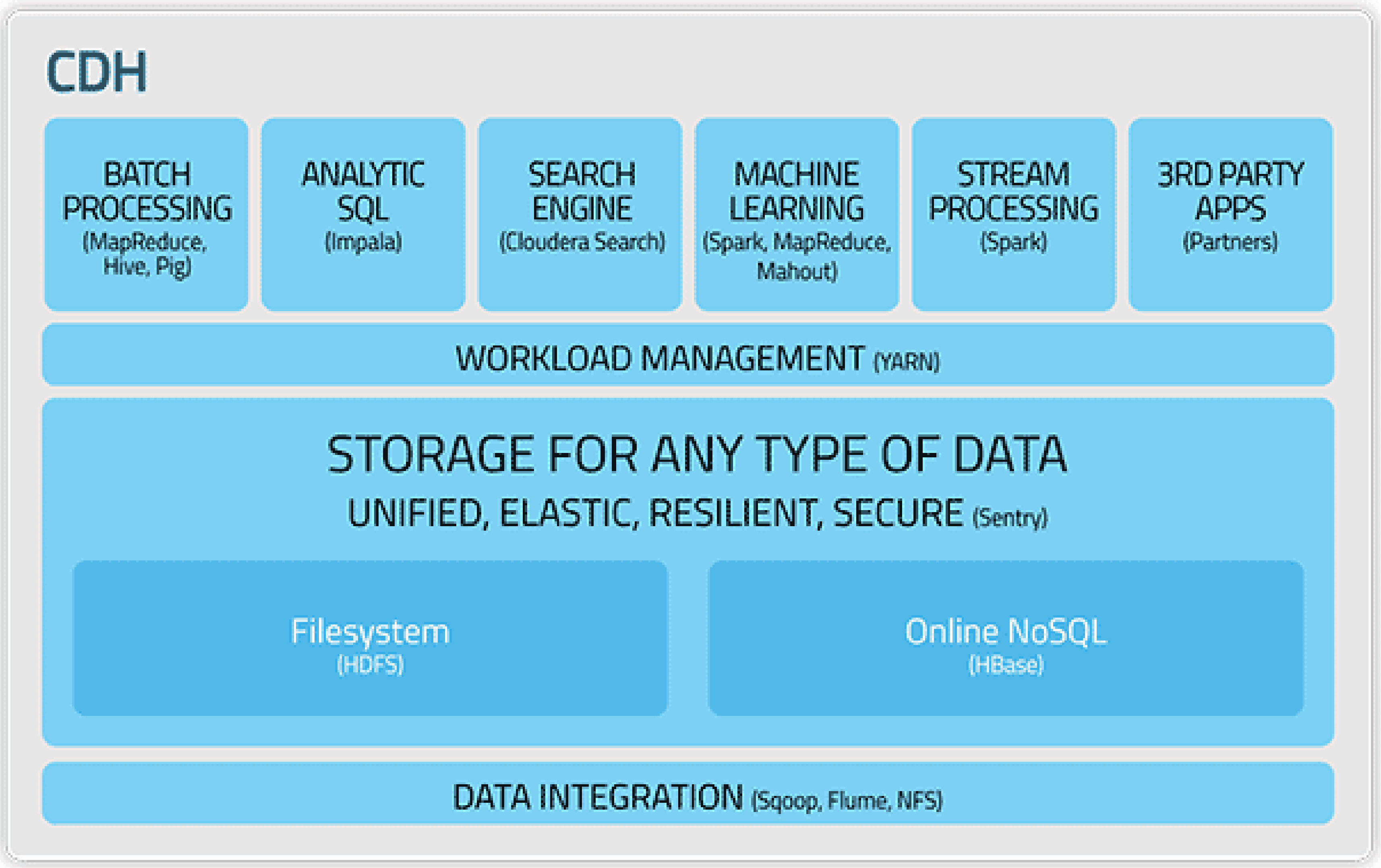
Mouse over each box for more details.

Click on each technology for resources to get started.

Last updated Oct 26th, 2015
You can also find the previous version here.

All these libraries/frameworks/components are designed and implemented to be distributed and can talk seamlessly to others at different layers of the ecosystem


OPERATIONAL SERVICES
AMBARI
FALCON*
OOZIE
DATA SERVICES
FLUME
SQOOP
LOAD & EXTRACT
NFS
WebHDFS
HBASE
PIG
HIVE & HCATALOG
MAP REDUCE
TEZ
YARN
CORE SERVICES
KNOX*
HDFS 2
Enterprise Readiness
High Availability, Disaster Recovery, Security and Snapshots
Hortonworks
HORTONWORKS DATA PLATFORM



CDH

BATCH PROCESSING (MapReduce, Hive, Pig)

ANALYTIC SQL (Impala)

SEARCH ENGINE (Cloudera Search)

MACHINE LEARNING (Spark, MapReduce, Mahout)

STREAM PROCESSING (Spark)

3RD PARTY APPS (Partners)

WORKLOAD MANAGEMENT (YARN)

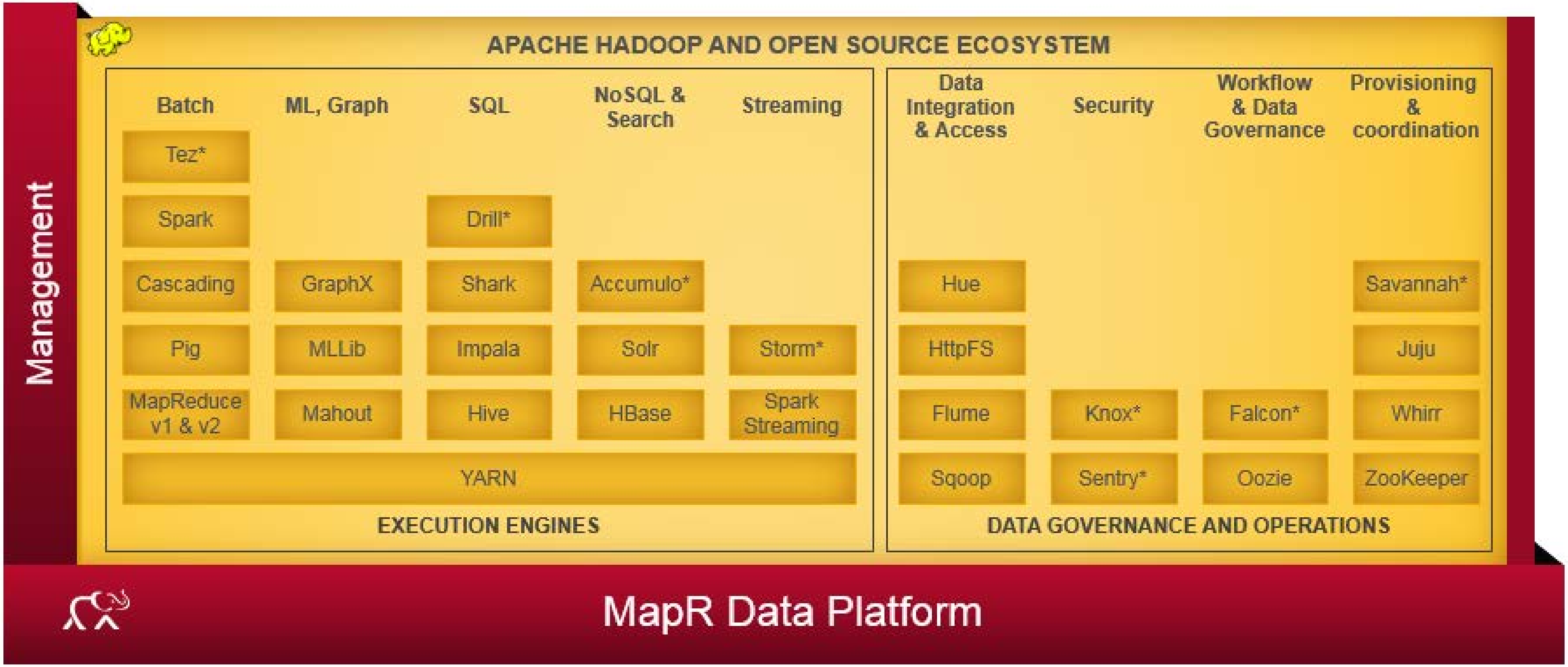
STORAGE FOR ANY TYPE OF DATA

UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

Filesystem (HDFS)

Online NoSQL (HBase)

DATA INTEGRATION (Sqoop, Flume, NFS)



APACHE HADOOP AND OPEN SOURCE ECOSYSTEM

Batch	ML, Graph	SQL	NoSQL & Search	Streaming
Tez*				
Spark		Drill*		
Cascading	GraphX	Shark	Accumulo*	
Pig	MLLib	Impala	Solr	Storm*
MapReduce v1 & v2	Mahout	Hive	HBase	Spark Streaming
YARN				

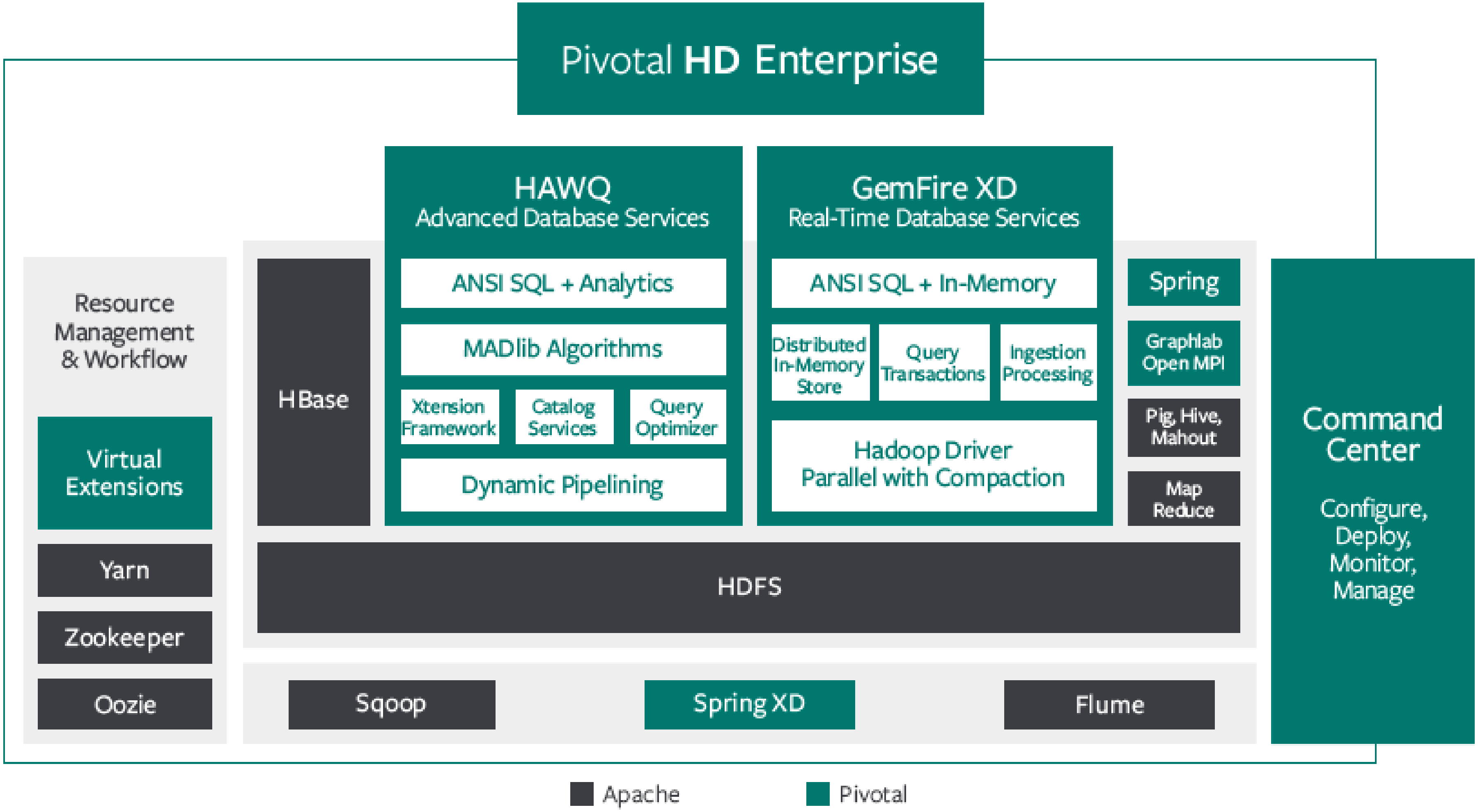
EXECUTION ENGINES

Data Integration & Access	Security	Workflow & Data Governance	Provisioning & coordination
Hue			Savannah*
HttpFS			Juju
Flume	Knox*	Falcon*	Whirr
Sqoop	Sentry*	Oozie	ZooKeeper

DATA GOVERNANCE AND OPERATIONS

Management

MapR Data Platform



Pivotal HD Enterprise

HAWQ
Advanced Database Services

ANSI SQL + Analytics

MADlib Algorithms

Xtension Framework

Catalog Services

Query Optimizer

Dynamic Pipelining

GemFire XD
Real-Time Database Services

ANSI SQL + In-Memory

Distributed In-Memory Store

Query Transactions

Ingestion Processing

Hadoop Driver
Parallel with Compaction

Resource Management & Workflow

Virtual Extensions

Yarn

Zookeeper

Oozie

HBase

Spring

Graphlab Open MPI

Pig, Hive, Mahout

Map Reduce

HDFS

Sqoop

Spring XD

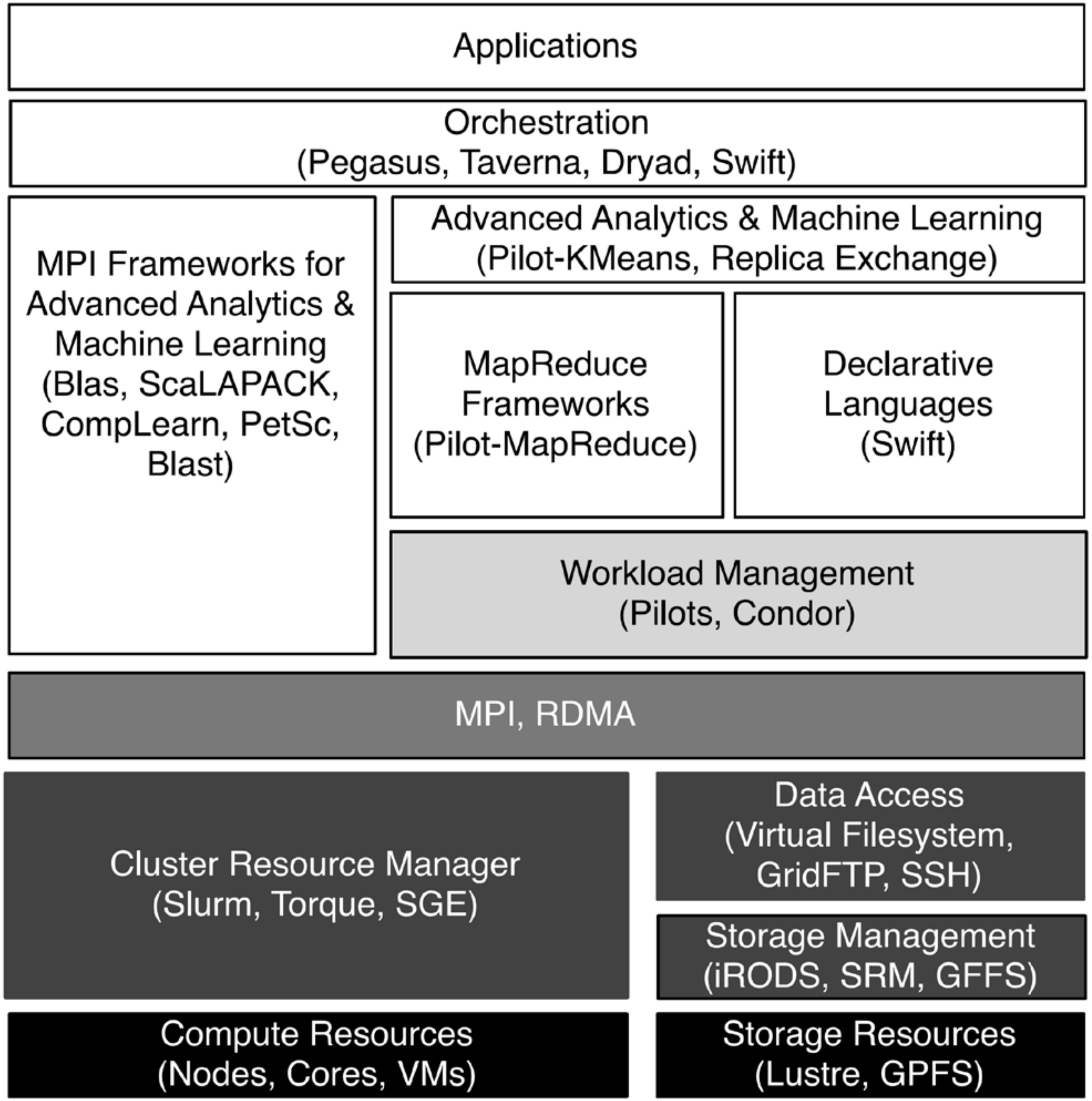
Flume

Command Center

Configure, Deploy, Monitor, Manage

■ Apache
🟩 Pivotal

How can we differentiate between a traditional high performance computing infrastructure (e.g., Beowulf/Palmetto) and a data-intensive computing infrastructure (e.g., Hadoop)



Applications

Orchestration
(Pegasus, Taverna, Dryad, Swift)

MPI Frameworks for Advanced Analytics & Machine Learning (Blas, ScaLAPACK, CompLearn, PetSc, Blast)

Advanced Analytics & Machine Learning
(Pilot-KMeans, Replica Exchange)

MapReduce Frameworks
(Pilot-MapReduce)

Declarative Languages
(Swift)

Workload Management
(Pilots, Condor)

MPI, RDMA

Cluster Resource Manager
(Slurm, Torque, SGE)

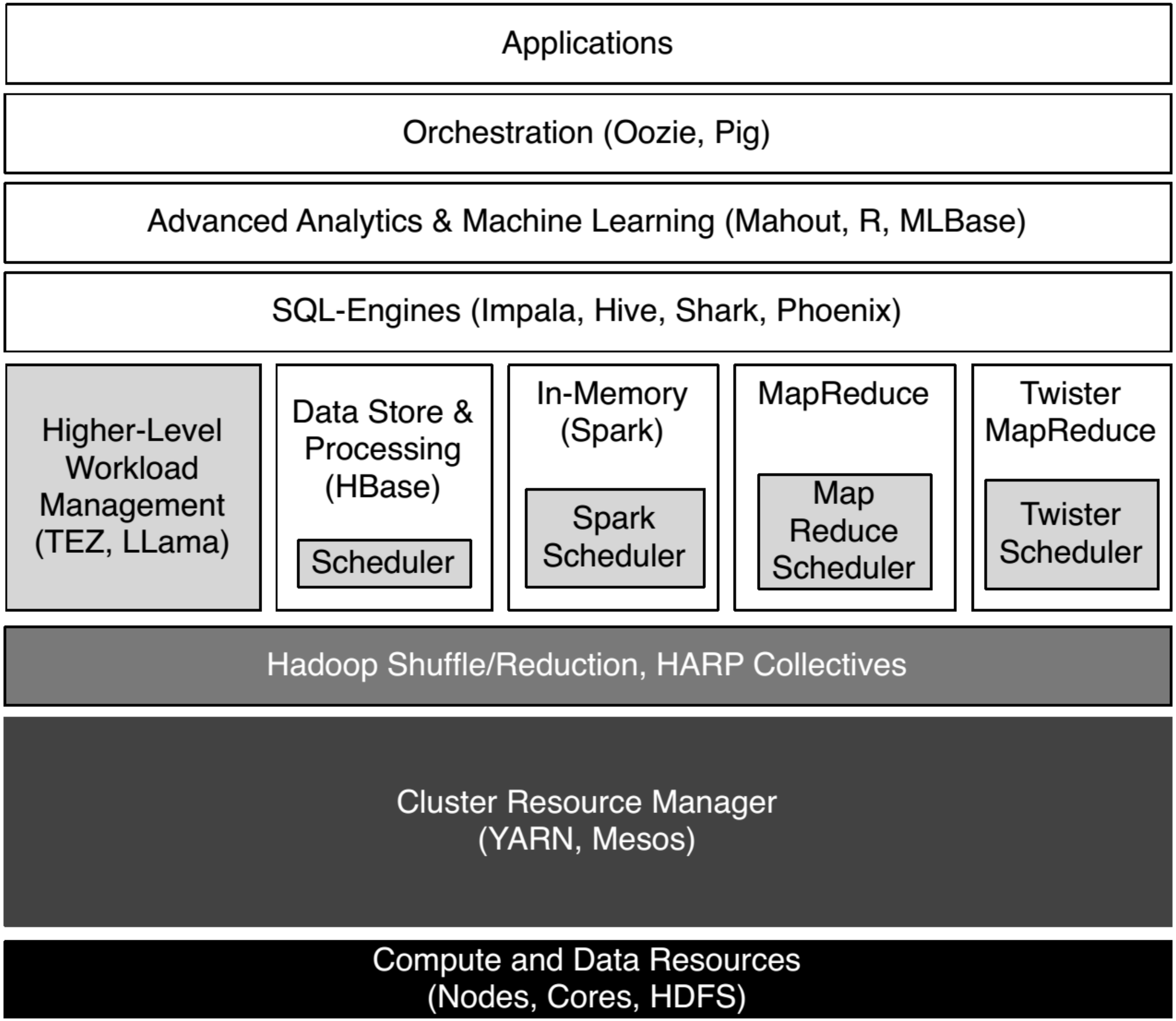
Data Access
(Virtual Filesystem, GridFTP, SSH)

Storage Management
(iRODS, SRM, GFFS)

Compute Resources
(Nodes, Cores, VMs)

Storage Resources
(Lustre, GPFS)

High-Performance Computing



Applications

Orchestration (Oozie, Pig)

Advanced Analytics & Machine Learning (Mahout, R, MLBase)

SQL-Engines (Impala, Hive, Shark, Phoenix)

Higher-Level Workload Management (TEZ, LLama)

Data Store & Processing (HBase)

Scheduler

In-Memory (Spark)

Spark Scheduler

MapReduce

Map Reduce Scheduler

Twister MapReduce

Twister Scheduler

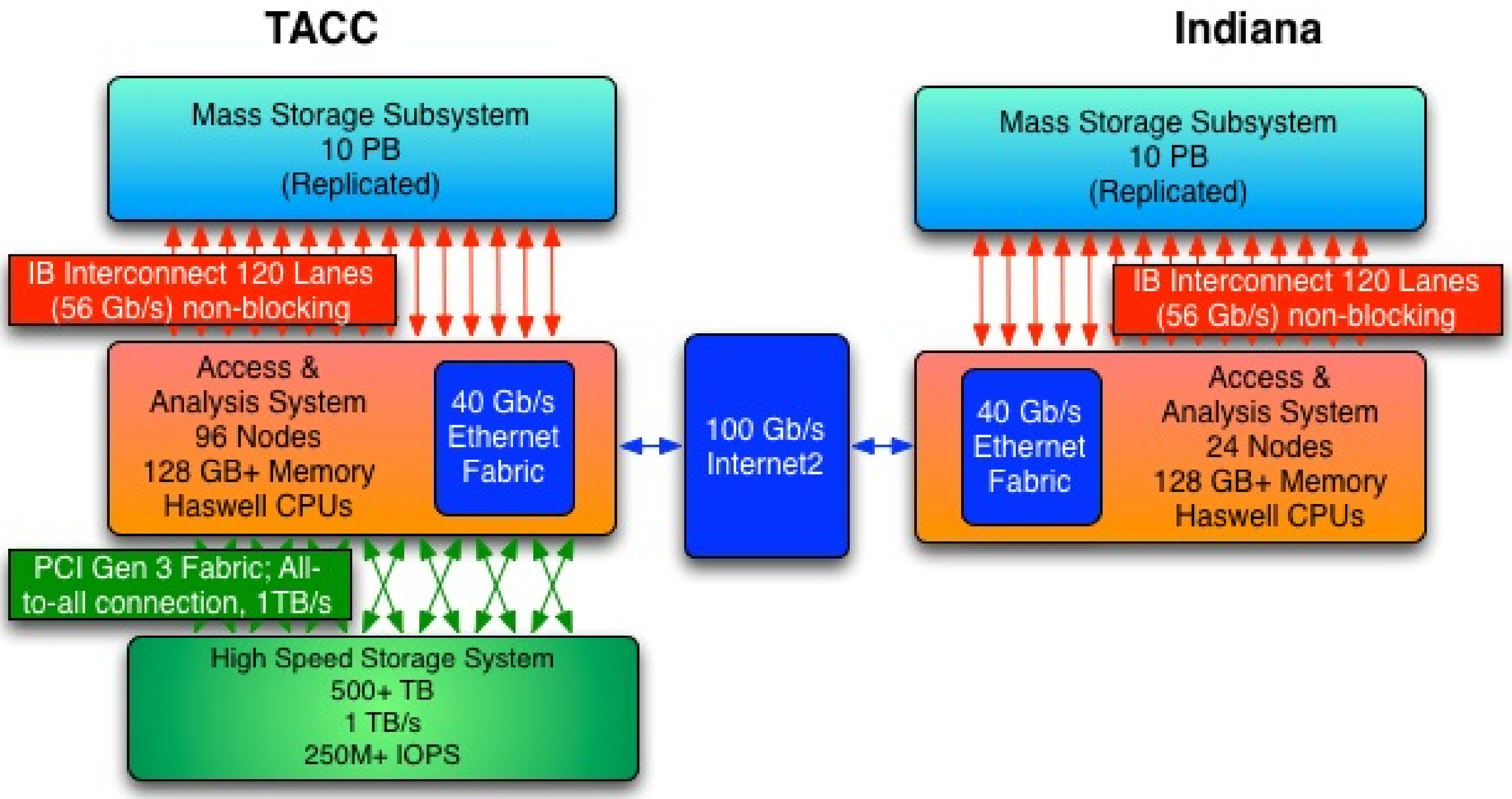
Hadoop Shuffle/Reduction, HARP Collectives

Cluster Resource Manager (YARN, Mesos)

Compute and Data Resources (Nodes, Cores, HDFS)

Apache Hadoop Big Data

Nextgen Computing Center



TACC

Mass Storage Subsystem
10 PB
(Replicated)

IB Interconnect 120 Lanes
(56 Gb/s) non-blocking

Access &
Analysis System
96 Nodes
128 GB+ Memory
Haswell CPUs

40 Gb/s
Ethernet
Fabric

PCI Gen 3 Fabric; All-to-all connection, 1TB/s

High Speed Storage System
500+ TB
1 TB/s
250M+ IOPS

100 Gb/s
Internet2

Indiana

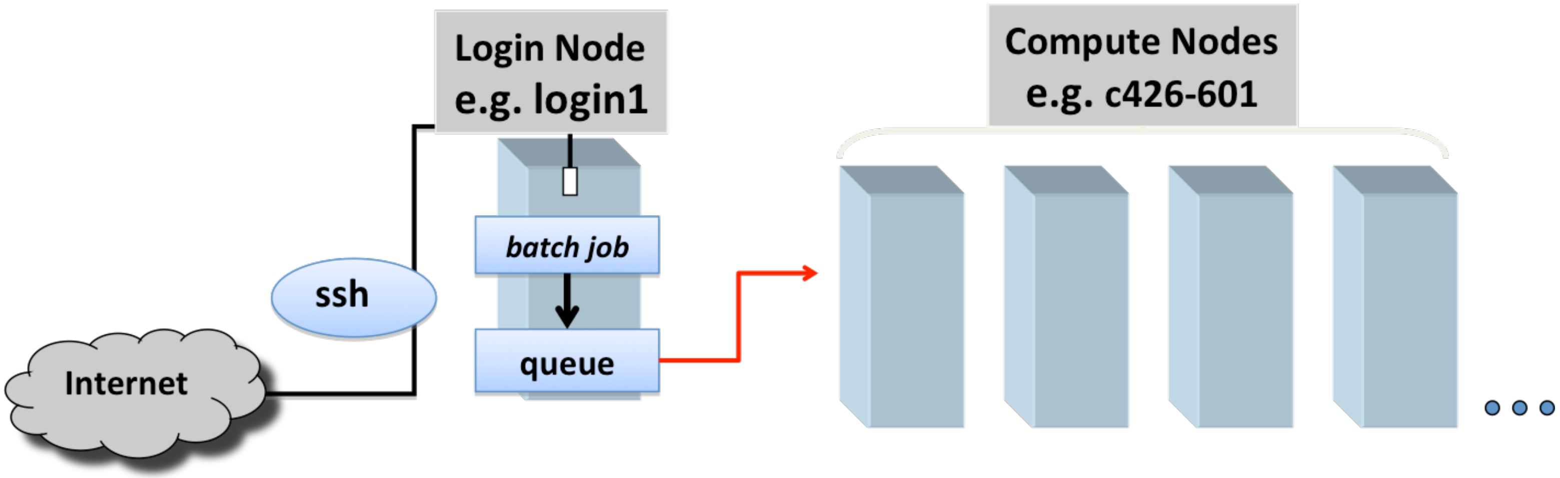
Mass Storage Subsystem
10 PB
(Replicated)

IB Interconnect 120 Lanes
(56 Gb/s) non-blocking

40 Gb/s
Ethernet
Fabric

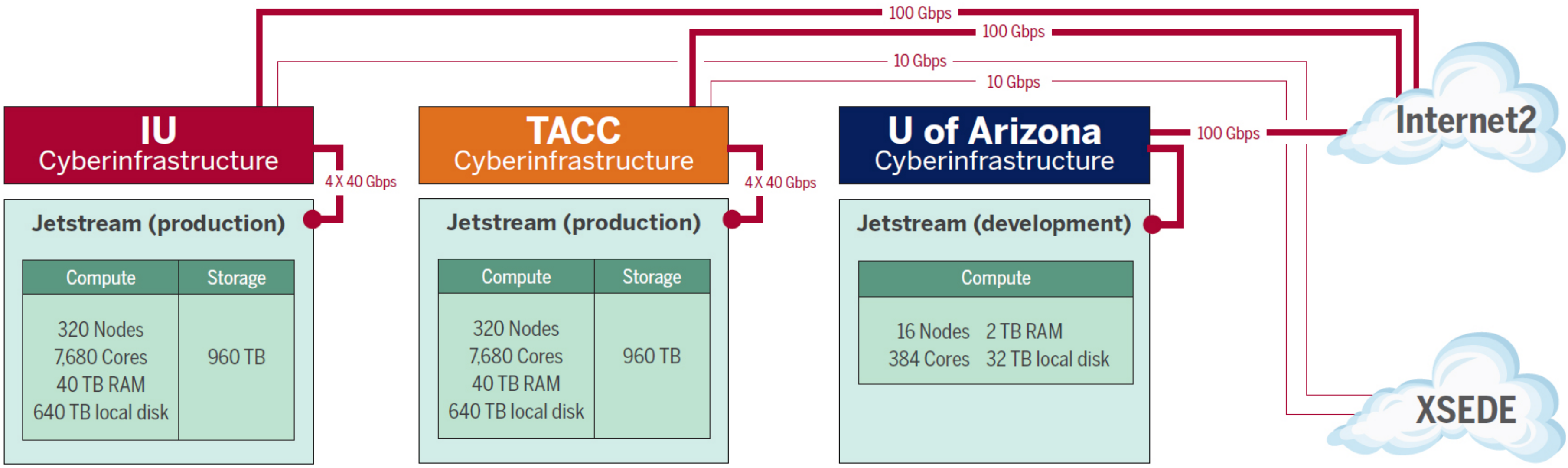
Access &
Analysis System
24 Nodes
128 GB+ Memory
Haswell CPUs

TACC Stampede (previous cluster)



Storage Class	Size	Architecture
Local	250GB/600GB/1TB	SATA
Parallel	14PB	Lustre
Tape	60PB	SAM-FS

Regional/National Resources



IU
Cyberinfrastructure

Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

TACC
Cyberinfrastructure

Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

U of Arizona
Cyberinfrastructure

Jetstream (development)

Compute	
16 Nodes 384 Cores	2 TB RAM 32 TB local disk

References

http://insightdataengineering.com/blog/new-ecosystem/

Hortonworks, 2015

Cloudera, 2015

MapR, 2015

Pivotal, 2015

Jha, Somesh, Jian Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox. "A tale of two data-intensive paradigms: Applications, abstractions, and architectures." In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pp. 645-652. IEEE, 2014.

https://portal.wrangler.tacc.utexas.edu/
https://portal.tacc.utexas.edu/user-guides/stampede#overview-table4