Introduction to Hadoop ecosystem

Do The Math

Chicago, IL
Bangalore, India
www.mu-sigma.com

March 15, 2012

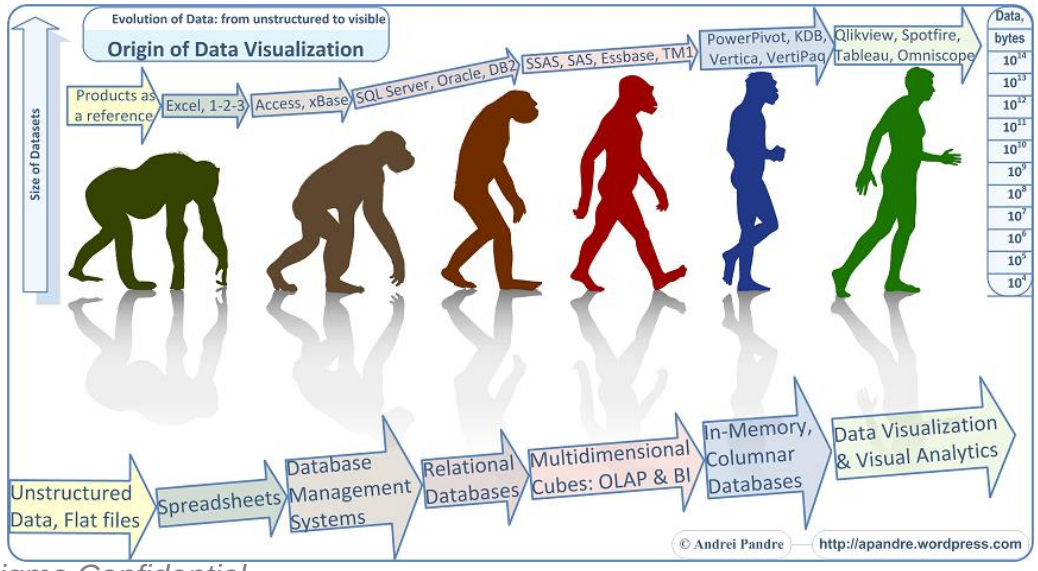
Agenda

- What is Big Data?
- Introduction to the Hadoop ecosystem
- Analytics using Hadoop, R, Mahout
- Competency requirements by focus area

What is Big Data?

- High Performance Computing
  - Parallel Computation
    - » “Beowulf” clusters, Grid Computing, Super Computing
  - Financial and Scientific Firms leading
    - » Real Time valuation, Monte-Carlo simulations
    - » Bioinformatics, Weather, Physical Sciences (CERN)

- Large Data Sets (**Volume, Velocity, Variability and Variety**)
  - Genesis was unstructured data (no consistent fields: web logs, machine logs, textual data, video, social data)
  - Ecommerce companies leading: Yahoo, Google, Facebook, Amazon (budget constraint: Map-Reduce!)



“**Big Data,**” claims GigaOm analyst Derrick Harris, is a bit of a misnomer; it’s really about data from different sources, including social networks and even cell phones. “It’s coming from sensors, it’s coming from computers, it’s coming from the Web,” he says.

Big data are datasets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing. This trend continues because of the benefits of working with larger and larger datasets allowing analysts to "spot business trends, prevent diseases, combat crime.”
source: Wikipedia

How Big is "Big Data"?

Kilobytes	Megabytes	Gigabytes	Terabytes	Petabytes	Exabytes	Zettabytes	Yottabytes
KB	MB	GB	TB	PB	EB	ZB	YB
=1024bytes	=1024kB	=1024MB	=1024GB	$=10^6$ GB	$=10^9$ GB	$=10^{12}$ GB	$=10^{15}$ GB

- Google processes **24 petabytes** of data per day (that's 24,000,000 GB)
  - *In 1993, total internet traffic was 100TB for the year*

- AT&T transfers **19 petabytes** of data on its network each day
  - *20 years of Hubble space telescope has amassed about 45 TB of data*

- The *Hippocampus* region of the brain, about an inch long, can store **2.4 petabytes** of binary data equivalent

- The movie Avatar took up **1 petabytes** of local storage for rendering of the 3D images
  - *In 2009, the 3D animated movie Monsters vs Aliens used 100TB of storage*

Agenda

- What is Big Data?
- Introduction to the Hadoop ecosystem
- Analytics using Hadoop, R, Mahout
- Competency requirements by focus area

Hadoop application areas and current challenges

What is hadoop ?

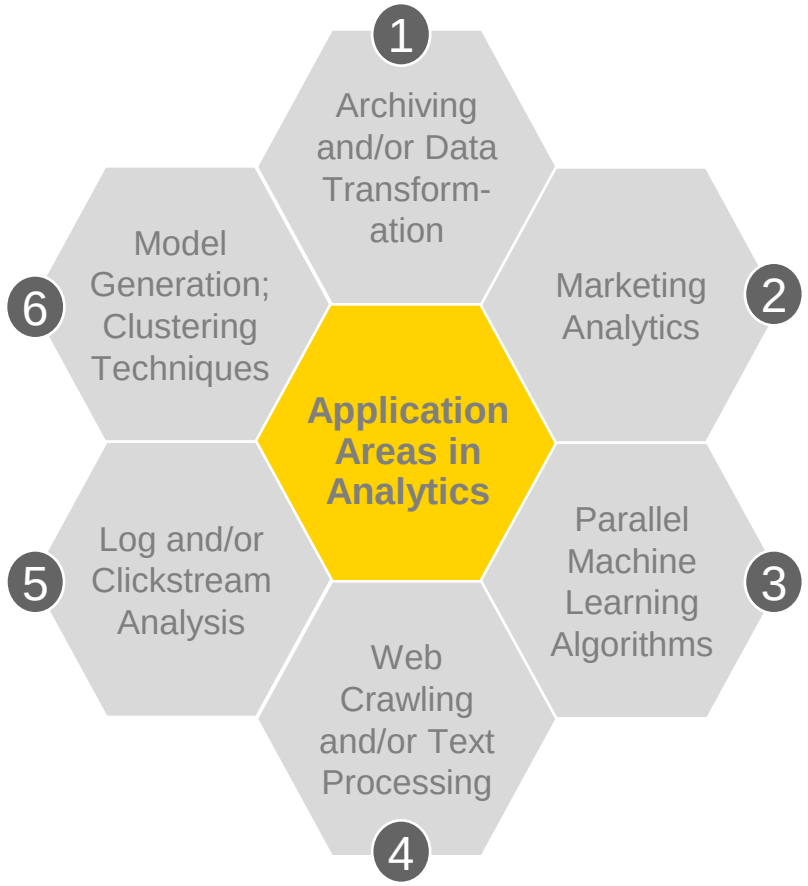
It is an open source project(Apache Software Foundation) for running applications on large clusters built of commodity hardware to process LARGE volumes of data.

Features:

Map/ Reduce	• Divides applications into fragments of work • Parallel distributed data processing paradigm
HDFS	• Distributed file system – files stored as blocks • Centralized management & reliability through replication

Current Analytical Challenges

- Data source explosion
- Technology complexities
- Computation and storage needs for expanding applications
- Advanced analytical techniques demanding parallelization



Simple, scalable, accessible and robust framework

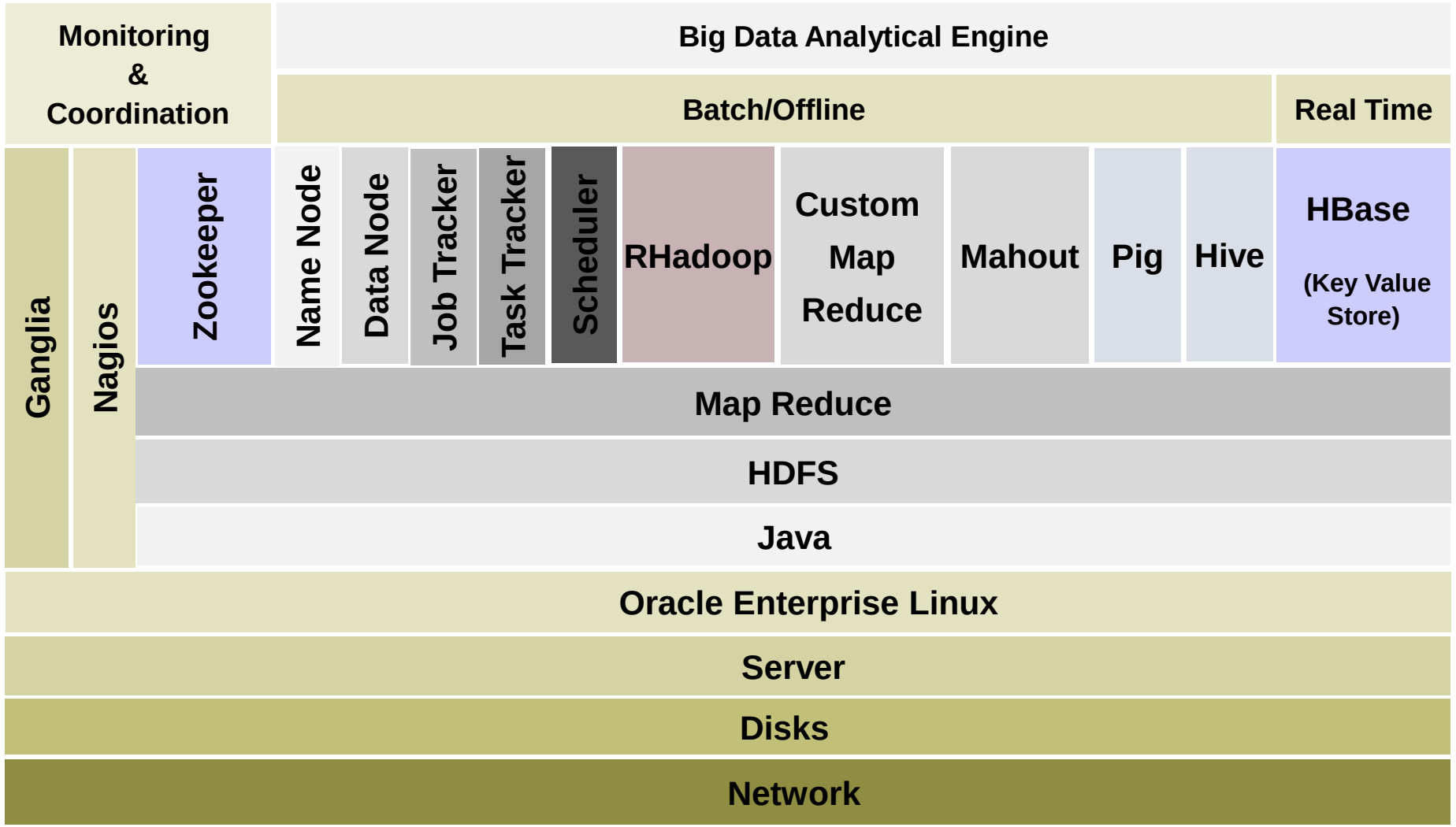
Example Usage…

LinkedIn	predict “People You May Know” and other facts
facebook	recommendation system, storage using HBASE
The New York Times	newspaper archive image conversion to PDF and text analytics
SPADAC	geospatial data indexing
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL	analyze gene sequence data
VISA	Credit card fraud analysis
YAHOO!	used to support AdSystems and web search
Journey Dynamics	analyze GPS records for traffic speed forecasting “To enable car insurance companies to assess the actual risk presented by an individual driver, based on driving behavior and driving patterns within the road context, and price that risk accordingly”

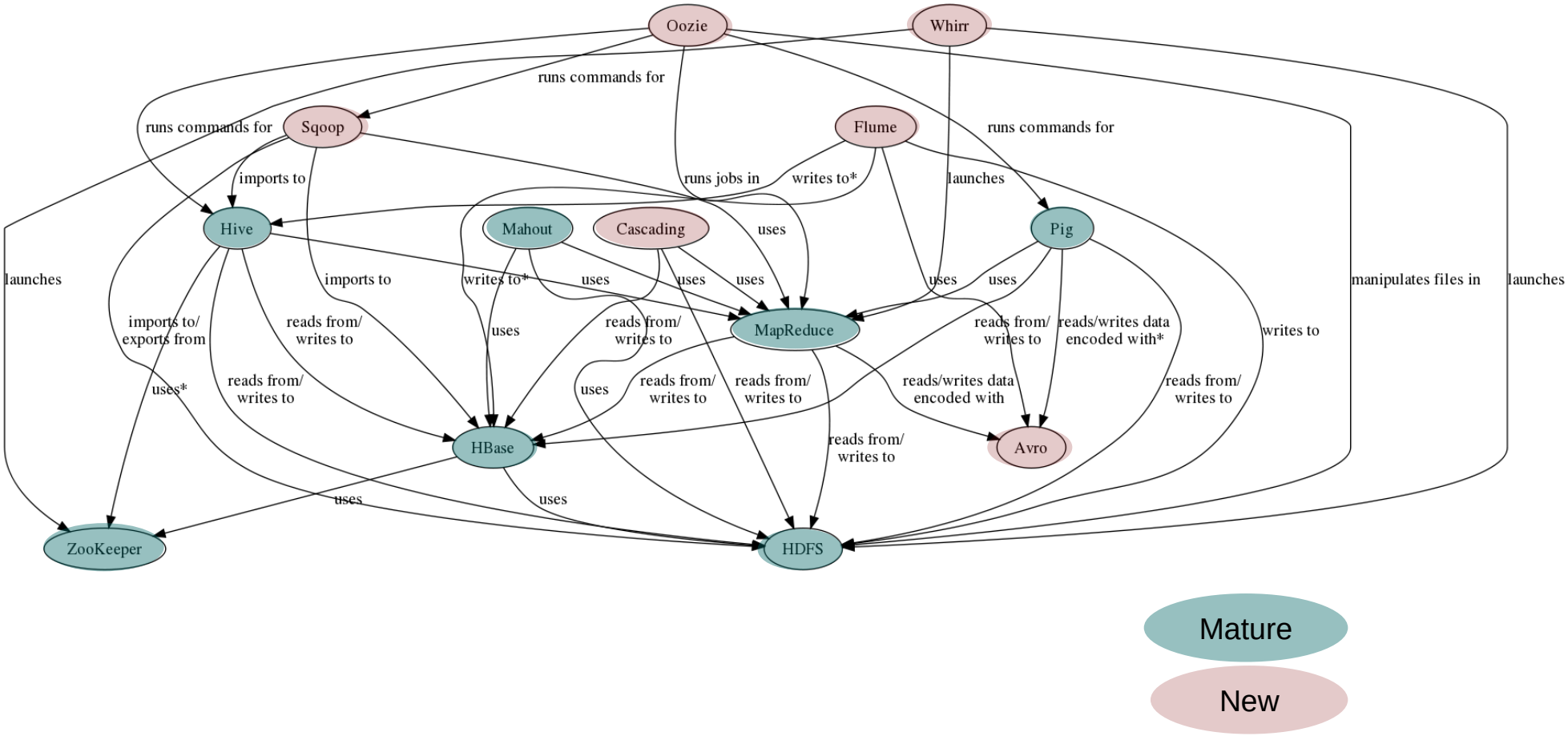
amazon ebay AOL twitter

*Source: IBM

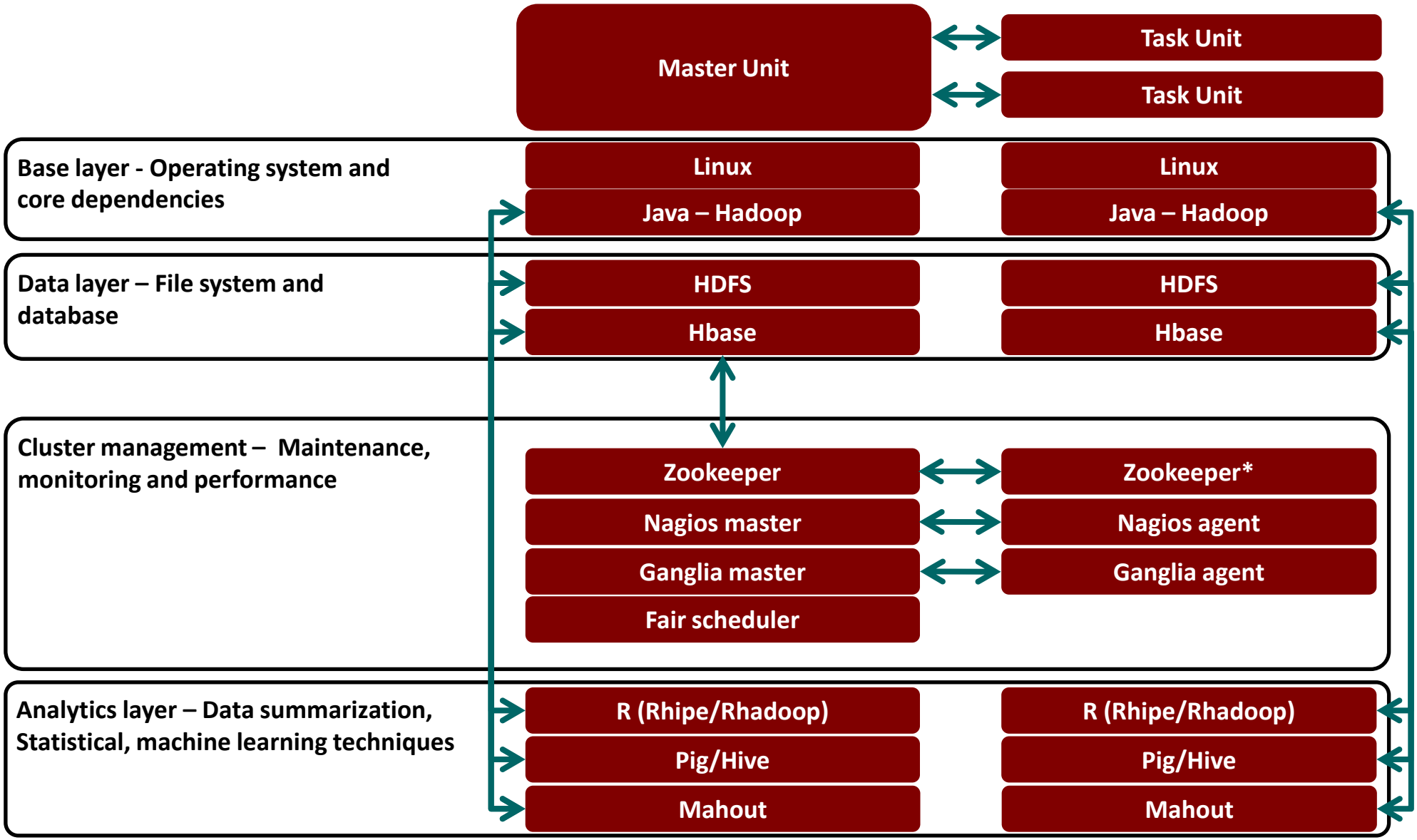
Hadoop Ecosystem



Hadoop is a complex ecosystem which is managed implicitly by HDFS (Hadoop Distributed File System)



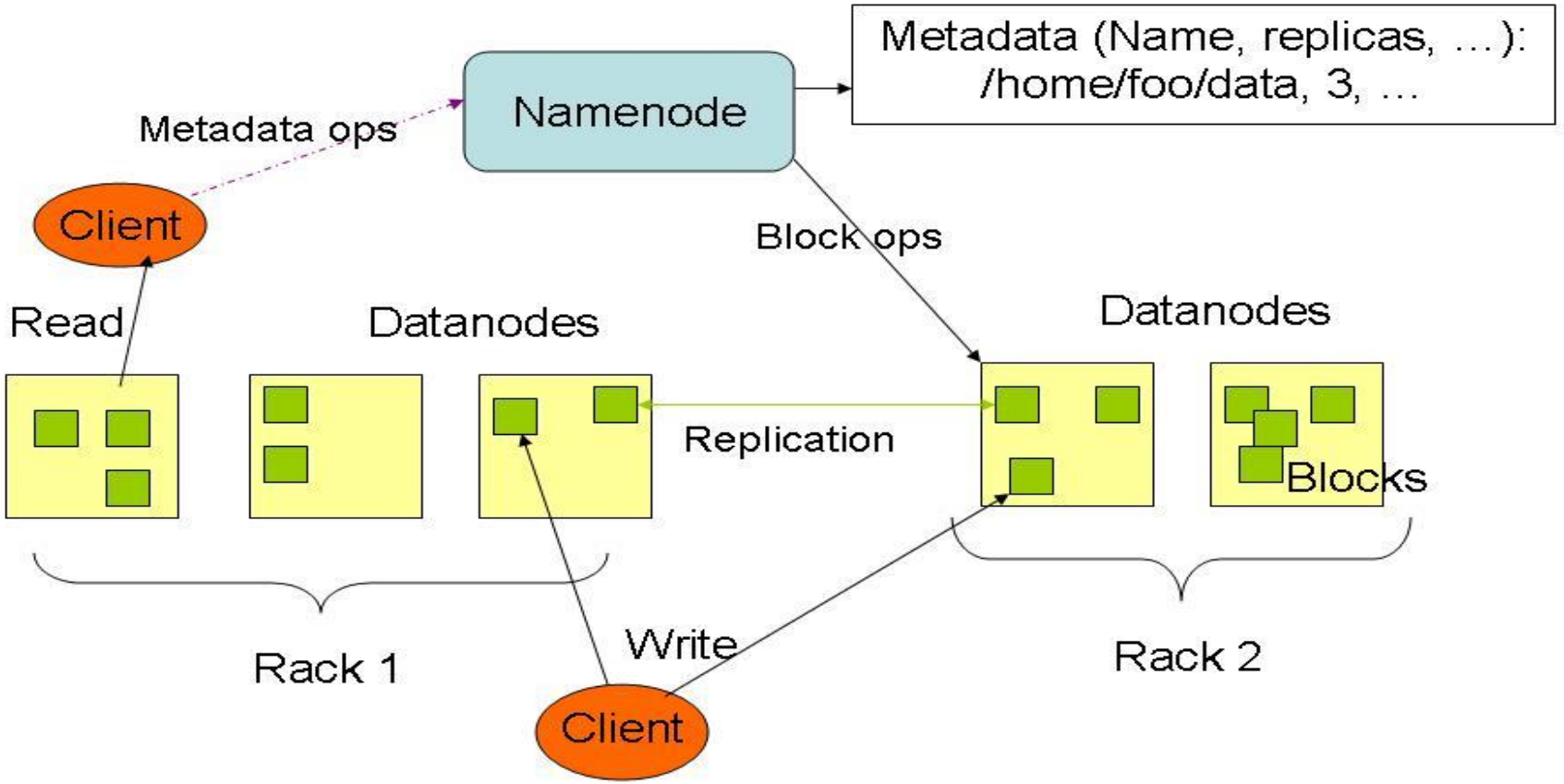
Overall Hadoop Architecture



** Zookeeper is installed by default along with HBASE; for DB operations, at least one Zookeeper instance, ideally on the name node, must be running at all times*

HDFS (Hadoop Distributed File Structure)

HDFS Architecture

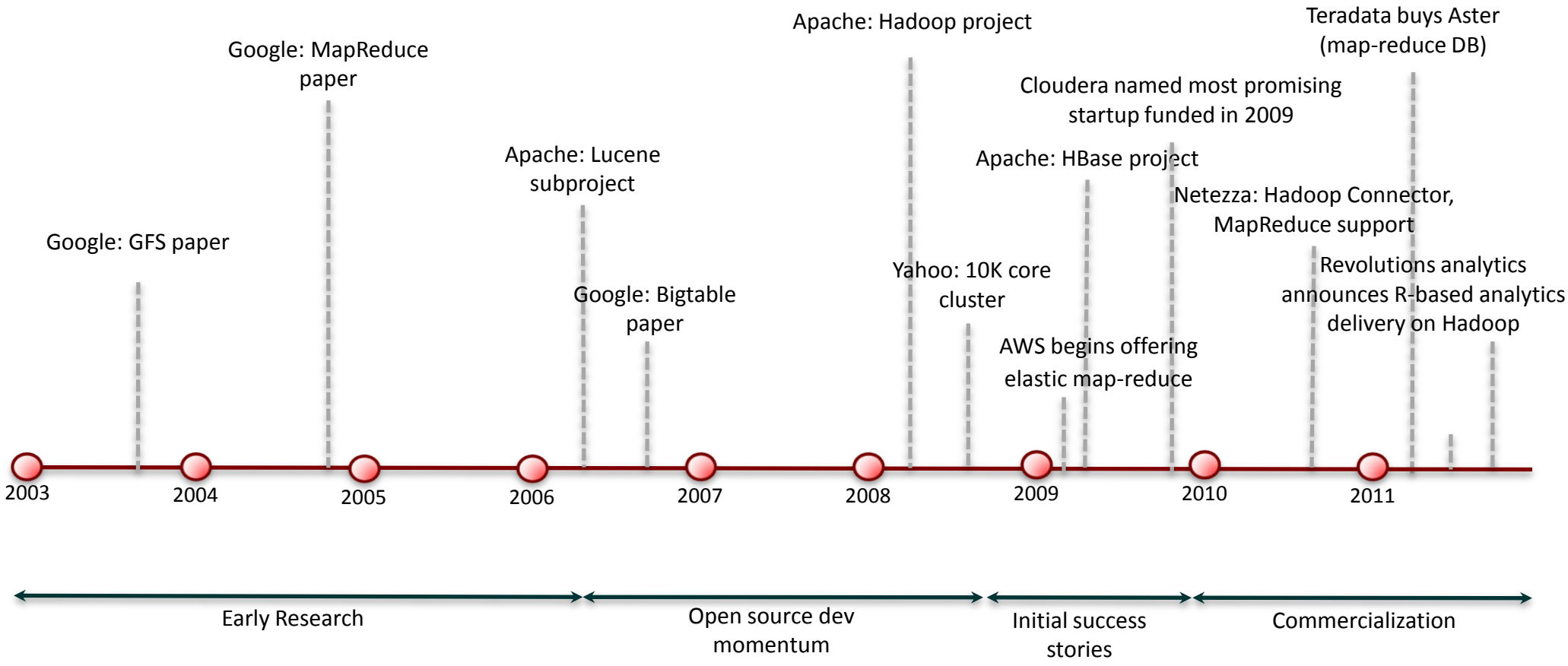


HDFS

- **Provide a single namespace for entire cluster**
  - Files, directories, and their hierarchy
- **Files are broken up into large blocks**
  - Typically 128 MB block size
  - Each block is replicated on multiple DataNodes
- **Meta-data in Memory**
  - Metadata: Names of files (including directories) and a list of Blocks for each file, list of DataNodes for each block, file attributes, e.g creation time, replication factor
  - High performance (high throughput, low latency)
- **A Transaction Log records file creations, file deletions etc**
- **Data Coherency:** emphasizes the append operation
- **Client can**
  - find location of blocks
  - access data directly from DataNode

Map-Reduce Concepts

Map-Reduce Origin and Evolution



Trend: Big Data (No SQL any language on Linux..)
Map Reduce Example – To Calculate Total Weekly Sales by Store

Raw Data → **Map Phase "Workers"** → **Reduce Phase**

Store	Weekly Sales ($)
A	1000
B	1500
C	2000
A	1200
B	1700
C	2200
A	2700
B	1700
C	2200

HDFS

Node 1

A	1000
B	1500
C	2000
A	1200
B	1700

A	(1000, 1200)
B	(1500, 1700)
C	2000

Node 2

C	2200
A	2700
B	1700
C	2200

A	2700
B	1700
C	(2200, 2200)

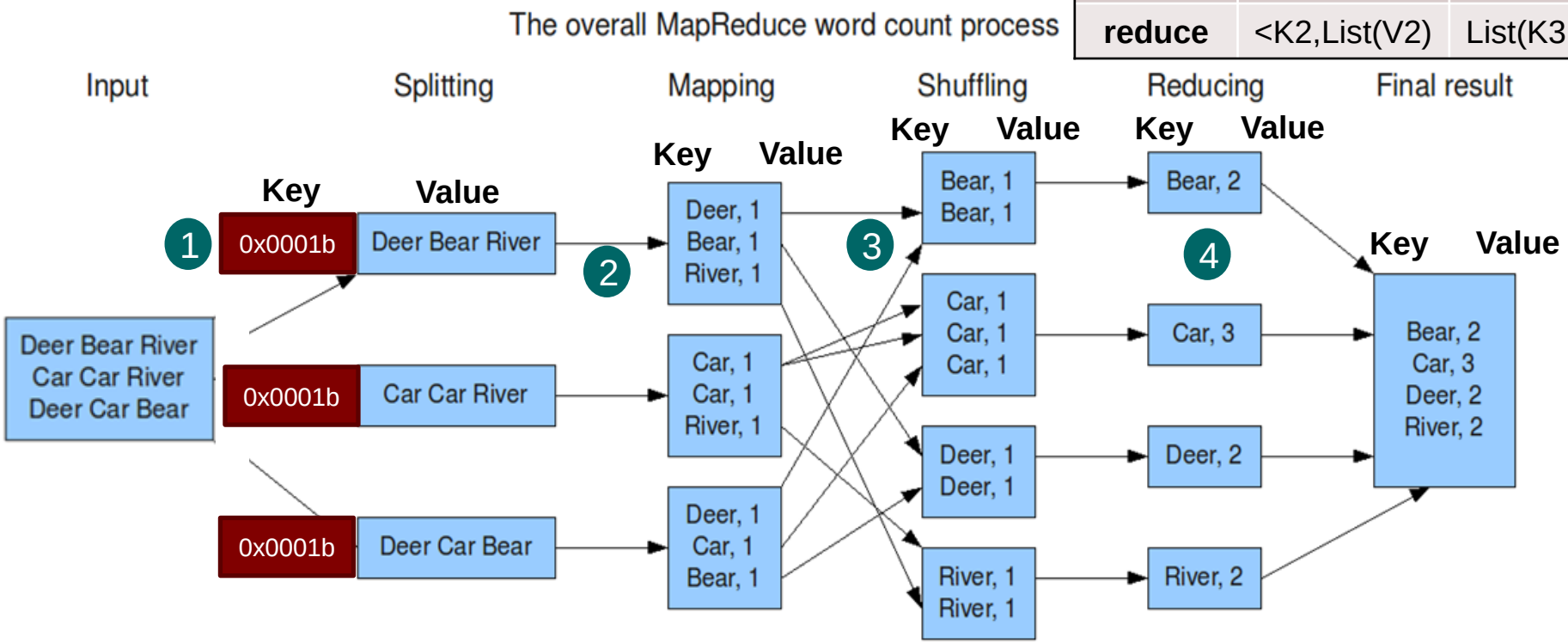
A	4900
B	4900
C	6400

"Reduce" step: The master node or set of reducers then takes the answers to all the sub-problems and combines them in some way to get the output – the answer to the problem it was originally trying to solve.

Maps are the individual tasks that transform input records into intermediate records.

Key-Value Example

	Input	Output
map	<K1,V1>	List(K2,V2)
reduce	<K2,List(V2)	List(K3,V3)



1. **By default, the line address of every line is they key and value is the contents of the line. These Key-Value pairs will be the input to the map function**

2. **Each Mapper will generate new Key-Value pairs based on the map function, In this case new key is the word and value is 1.**

3. **Key-Value pairs will be merged on the same key before sending it to the reducer. In this case, since word is the key, so all the key-value pairs associated with same word are merged**

4. **Finally, the reducer function run on the values associated to same key, and produces the result**

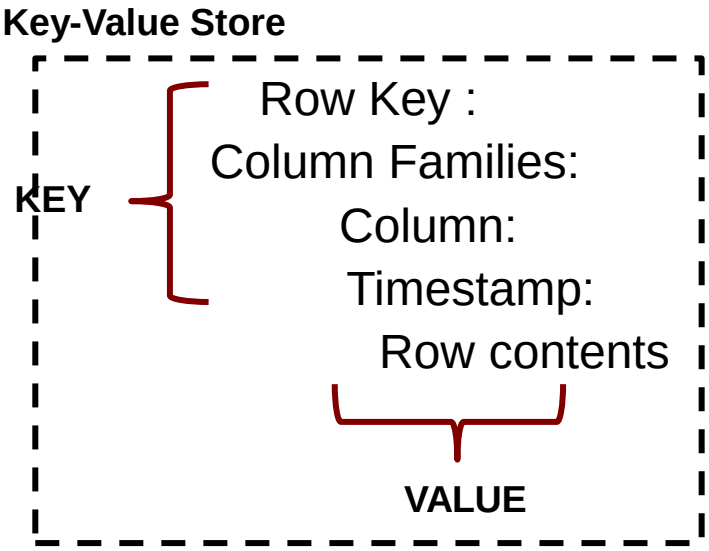
Apache HBase

Apache HBase

- HBase is the Hadoop database, which provides random, real-time read/write access to your Big Data, thus allows hosting of very large tables -- billions of rows X millions of columns on top of Hadoop cluster.

- **Clone of Google's BigTable**
  - Distributed (automatic partitioning)
  - Column-oriented
  - Semi-structured (columns can be added just by inserting)
  - Built-in versioning

- **Not an RDBMS**
  - No joins
  - No SQL
  - Data usually not normalized
  - Transactions & built-in secondary indexes available (as contrib.) but immature



- **Need to think differently about how you structure data**
  - De normalize your data where necessary
  - Structure data & row keys around common access

- HBase Use cases : http://wiki.apache.org/hadoop/Hbase/PoweredBy

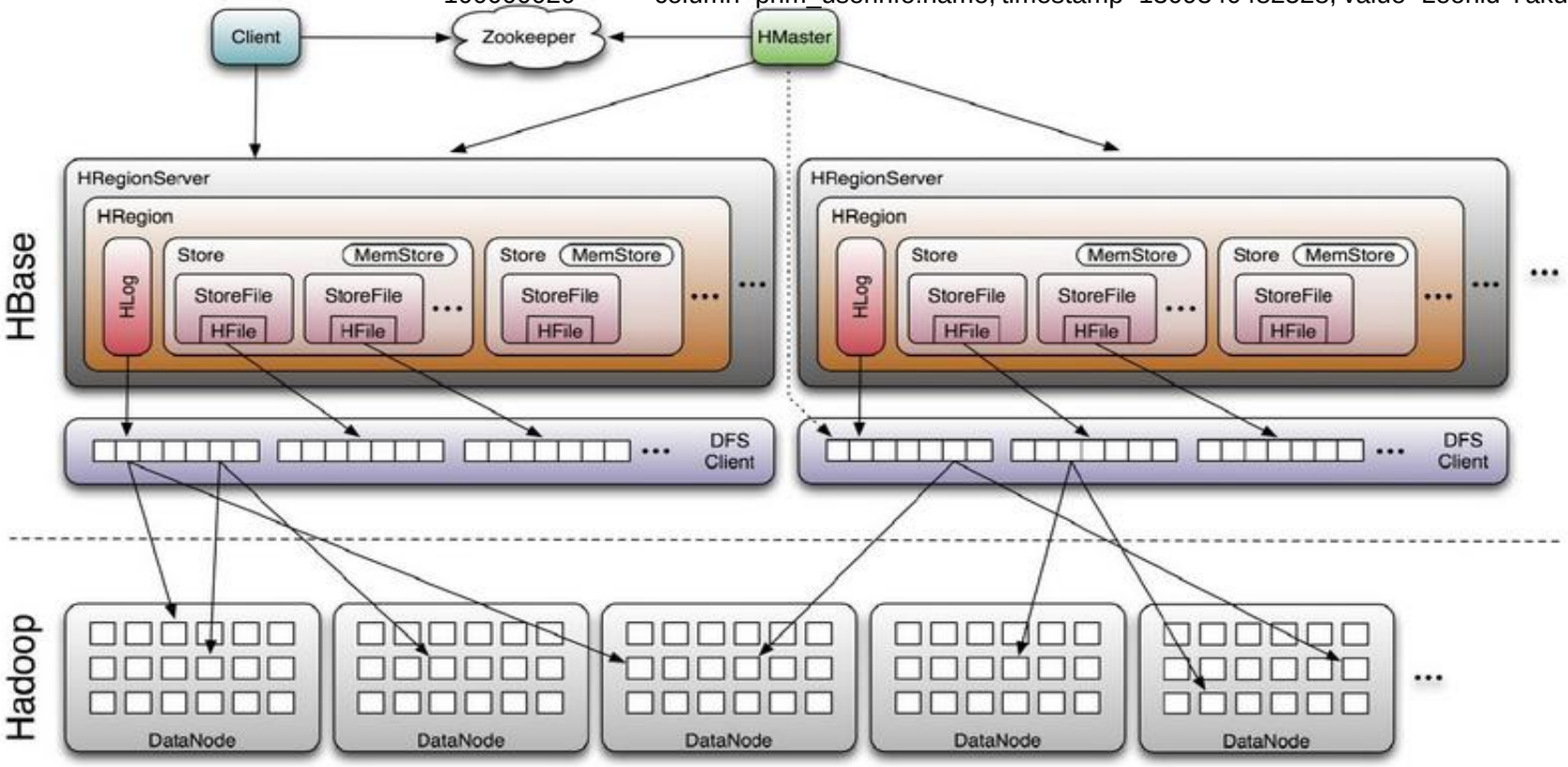
HBase Architecture

HBase Sample Data :

RowKey Column = Column Family: Column Name, timestamp, Row value

hbase(main):003:0> scan 'twitter_userinfo', LIMIT=>5

ROW	COLUMN+CELL
100000020	column=prim_userinfo:favcount, timestamp=1309340482528, value=0
100000020	column=prim_userinfo:followercount, timestamp=1309340482528, value=159
100000020	column=prim_userinfo:friendcount, timestamp=1309340482528, value=200
100000020	column=prim_userinfo:location, timestamp=1309340482528, value=null
100000020	column=prim_userinfo:name, timestamp=1309340482528, value=Leonid Yakub



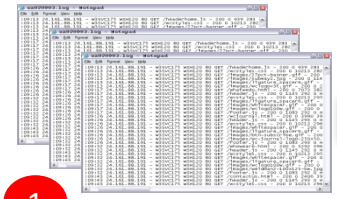
Hadoop and RDBMS Coexistence

unstructured data
or analytics datasets

Create context (classification, text mining, statistics & machine learning)

RDBMS: Analyze & Reporting

Data Sources



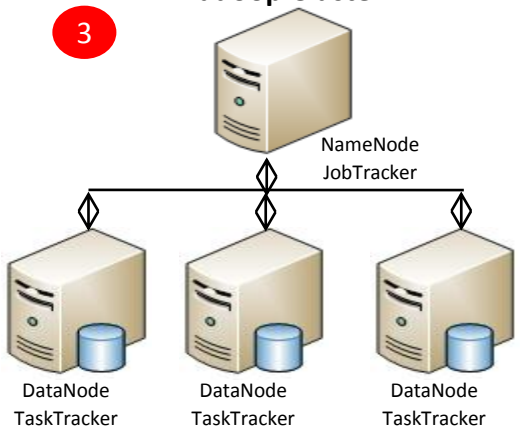
1

Raw Weblogs

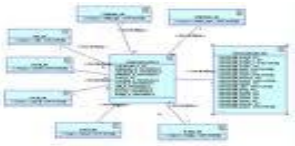
2

Hadoop Cluster

3



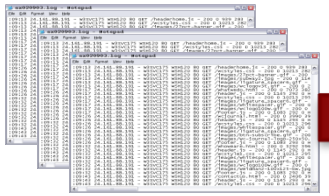
4



structured data

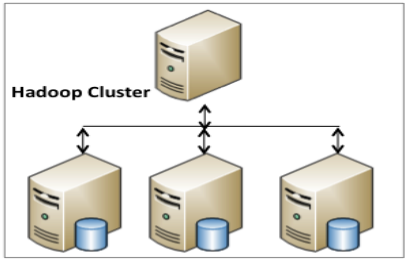
RDBMS: Analyze & Reporting

Active archival
Long running queries



1

2

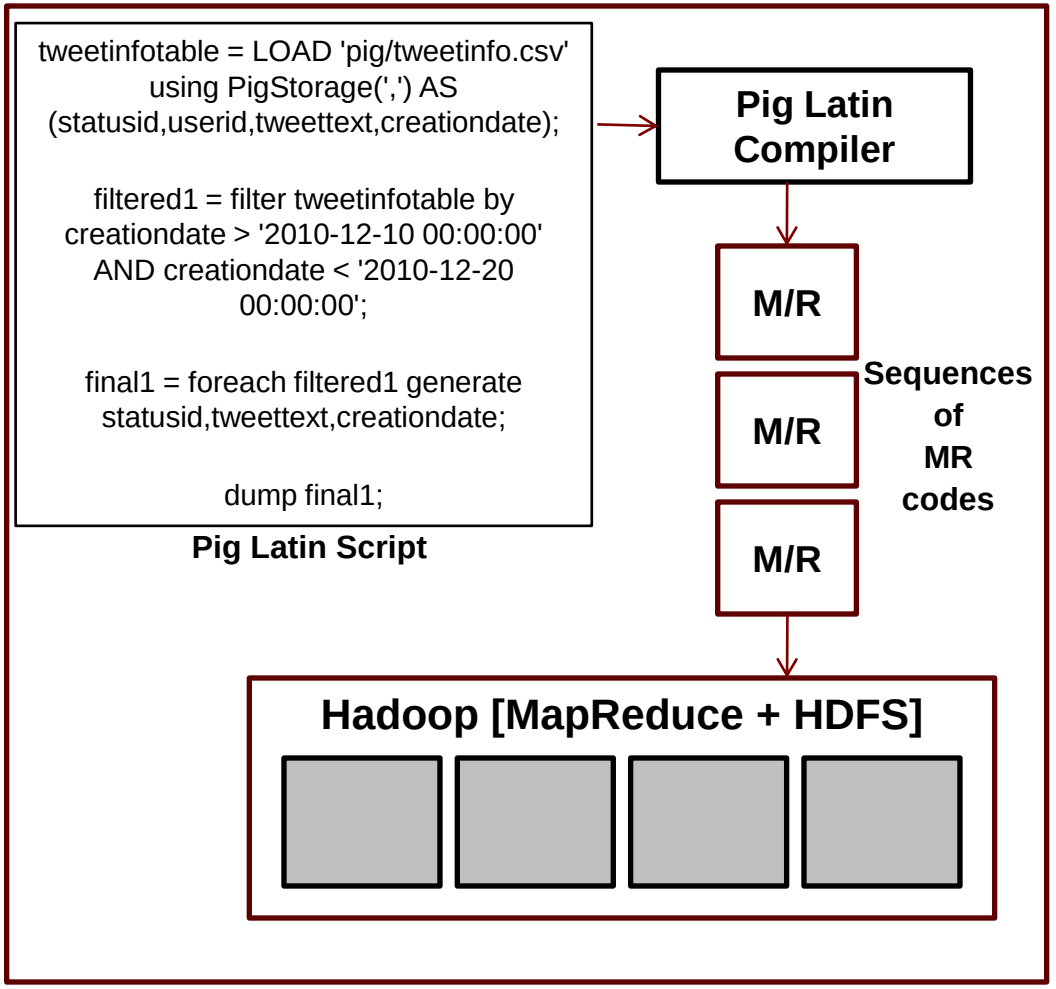


Example:
Pictures/Images
POS Receipts

Apache Pig

Apache Pig

- Apache Pig, developed by Yahoo, is a platform for analyzing large data sets that uses Hadoop map-reduce framework and HDFS.
- It provides an engine for executing data flows in parallel on Hadoop



Pig's infrastructure layer consists of

- a compiler that produces sequences of Map-Reduce programs,
- Pig's language layer currently consists of a textual language called **Pig Latin**
- includes operators for many of the traditional data operations (join, sort, filter, etc.) as well as the ability for users to develop their own functions for reading, processing, and writing data

Pig Philosophy

- **Pigs Eat Anything**
  - Pig can operate on data whether it has metadata or not. It can operate on data that is relational, nested, or unstructured.
- **Pigs Live Anywhere**
  - Pig is intended to be a language for parallel data processing. It is not tied to one particular parallel framework.
- **Pigs Are Domestic Animals**
  - Pig is designed to be easily controlled and modified by its users. Pig allows integration of user code where ever possible, so it currently supports user defined field transformation functions, user defined aggregates, and user defined conditionals.
- It is a good alternative to do parallel data processing on Hadoop than writing your own personal MR codes. Your 100 lines of Java MR code may get reduced to simple 3-5 pig-Latin lines.

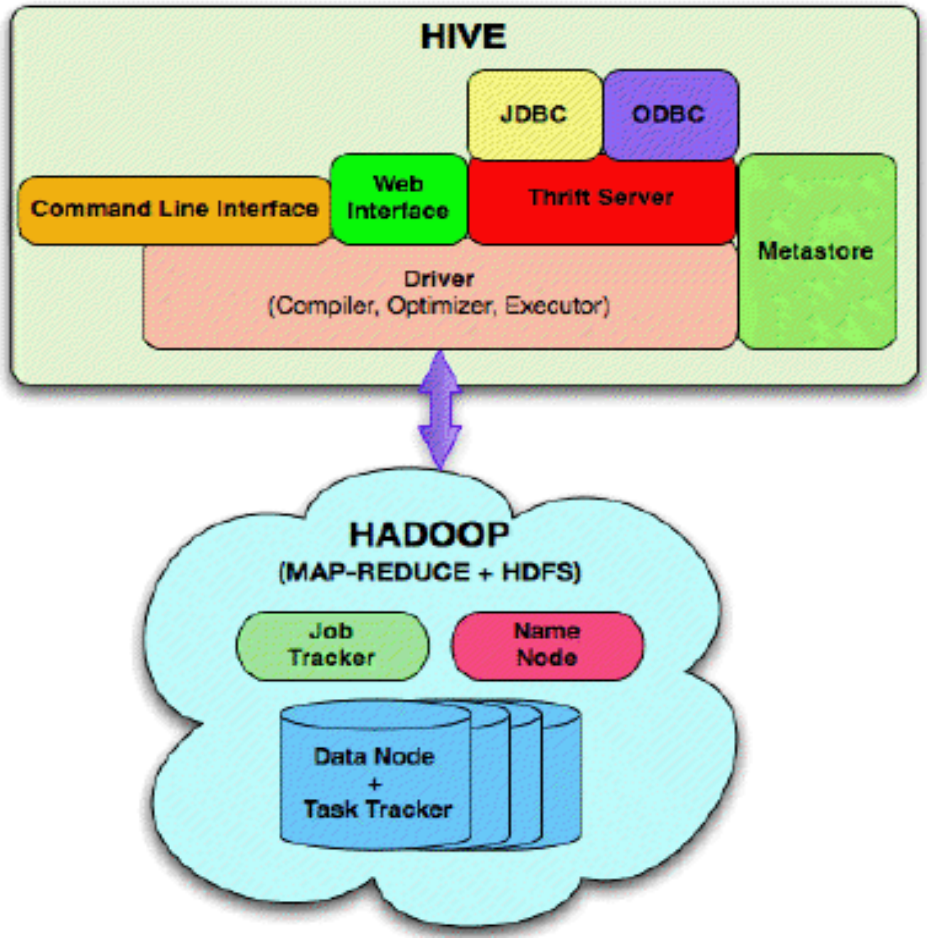
Apache Pig use-cases

- **Traditional extract transform load (ETL)**
- **Data pipelines**
  - A common example is web companies bringing in logs from their web servers, cleansing the data, and pre computing common aggregates before loading it into their data warehouse.
- **Research on raw data**
  - Since Pig can operate in situations where the schema is unknown or incomplete or inconsistent and since it can easily manage nested data, researchers who want to work on data before it has been cleaned and loaded into the warehouse often prefer Pig
- **Iterative processing**

Apache HIVE

Apache HIVE

- Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix.
- It provides an SQL-like language called HiveQL while maintaining full support for map/reduce.
- Internally, a compiler translates HiveQL statement into a directed acyclic graph of MapReduce jobs, which are submitted to Hadoop for execution
- **Use Cases**
  - *For data management, analysis, log aggregation, reporting, ETL into Hive*

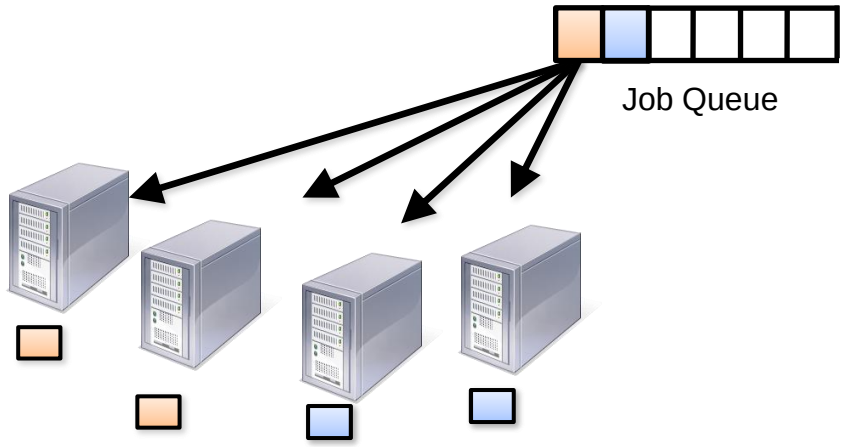


Pig Vs HIVE

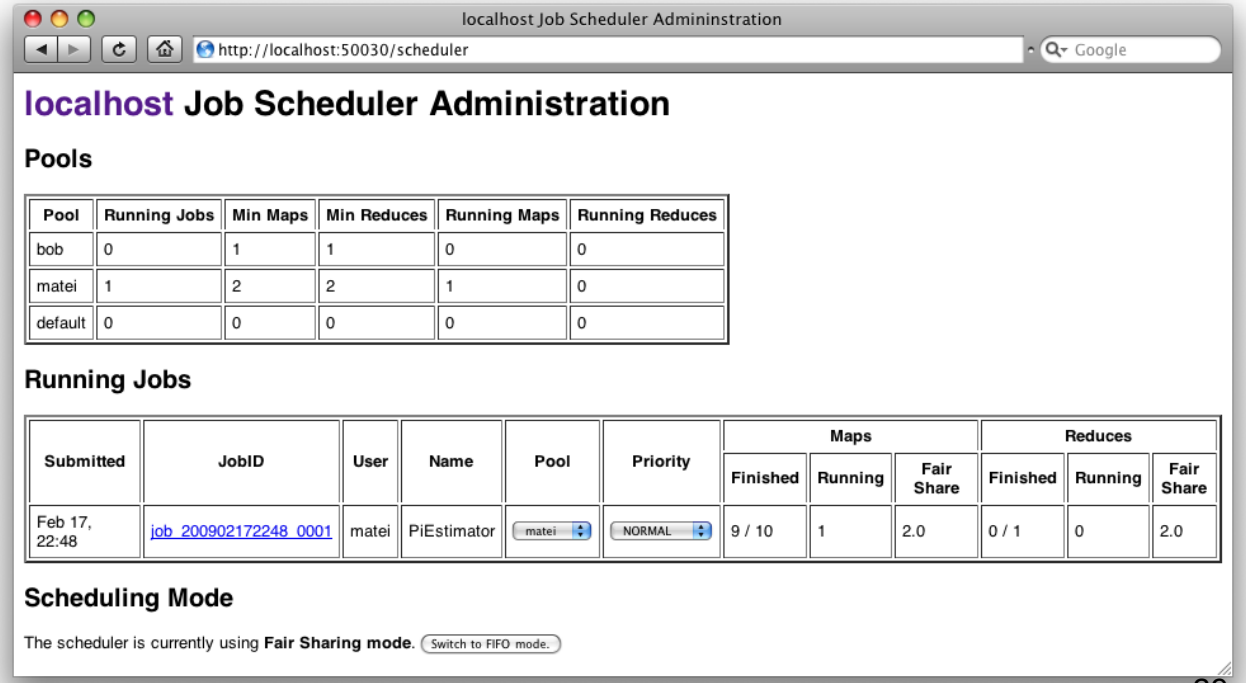
Feature	HIVE	Pig
Language	SQL-like	Pig Latin
Schemas/Types	Yes (explicit)	Yes (implicit)
Partitions	Yes	No
Server	Optional (Thrift)	No
User Defined Functions (UDF)	Yes (Java)	Yes (Java)
Custom Serializer/Deserializer	Yes	Yes
DFS Direct Access	Yes (implicit)	Yes (explicit)
Join/Order/Sort	Yes	Yes
Shell	Yes	Yes
Streaming	Yes	Yes
Web Interface	Yes	No
JDBC/ODBC	Yes (limited)	No

Scheduling and monitoring tools

Fair Scheduler

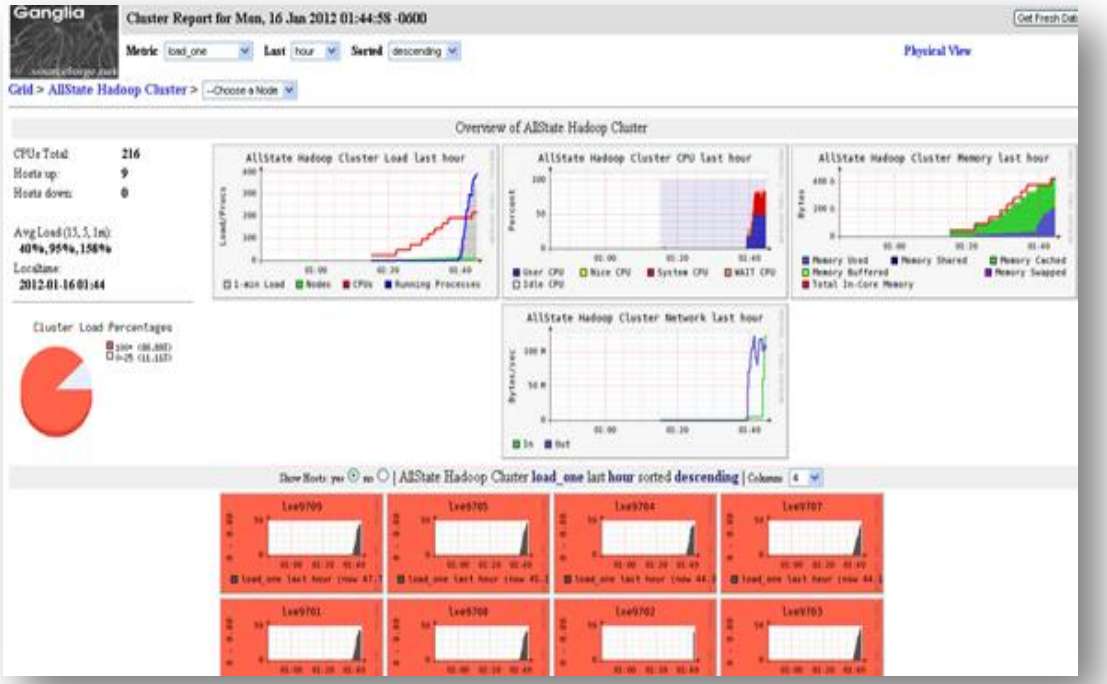


- Group jobs into "*pools*"
- Assign each pool a guaranteed *minimum share* (split up among its jobs)
- Split excess capacity evenly between jobs
- Limits on # of running jobs:
  - Per user
  - Per pool



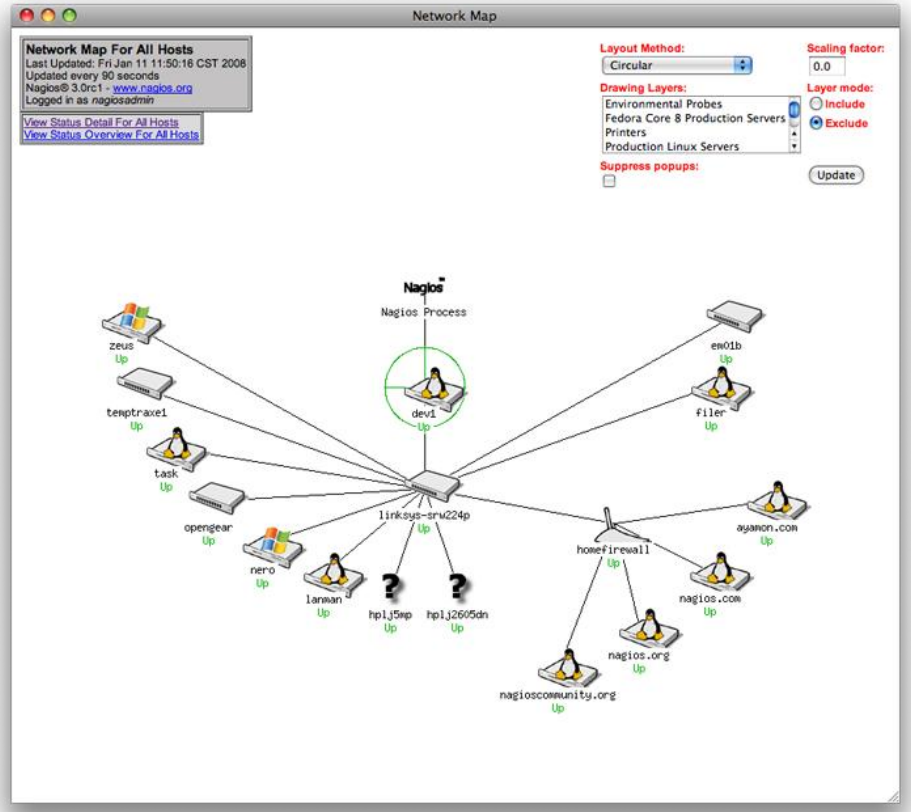
Ganglia

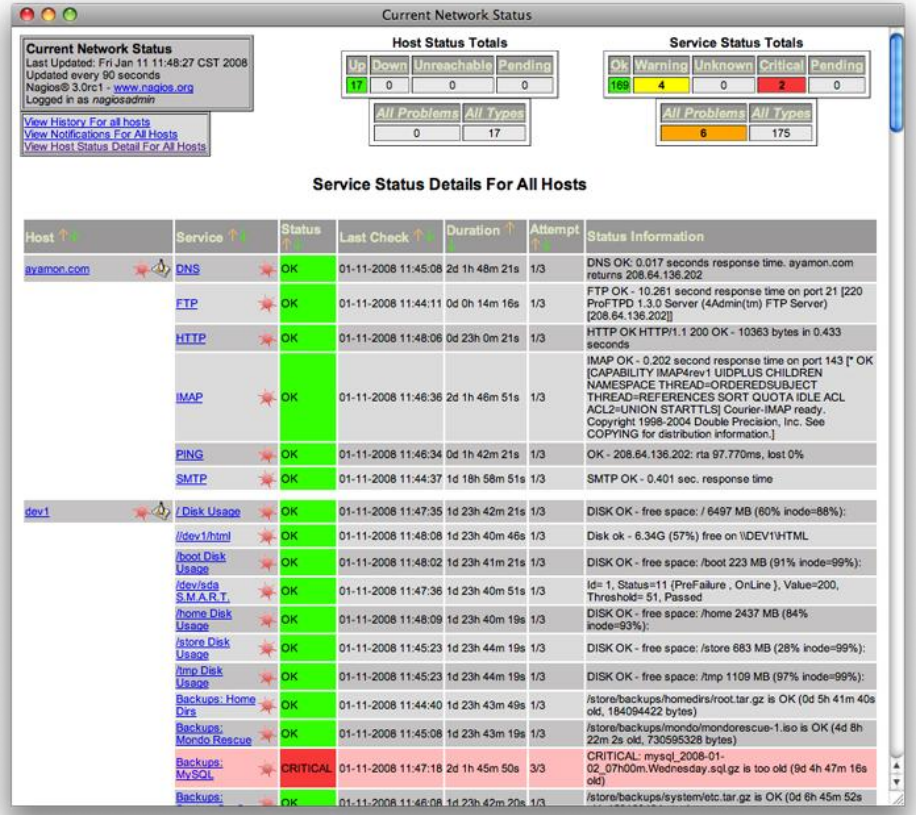
- Scalable distributed monitoring system for high-performance computing systems such as clusters and Grids
  - Heartbeat messages on a well-known multicast address enables automatic discovery of nodes. No manual configuration of cluster membership lists.
  - Each node monitors its local resources and lets others know of its state.
  - Each node listens to monitoring data from other nodes. Therefore, any node knows the entire state of the cluster.
- Web based User Interface
- A data collector and trending tool



Nagios

- Cluster System monitor
- No graphs, focused on service and host uptime monitoring.





Agenda

- What is Big Data?
- Introduction to the Hadoop ecosystem
- Analytics using Hadoop, R, Mahout
- Competency requirements by focus area

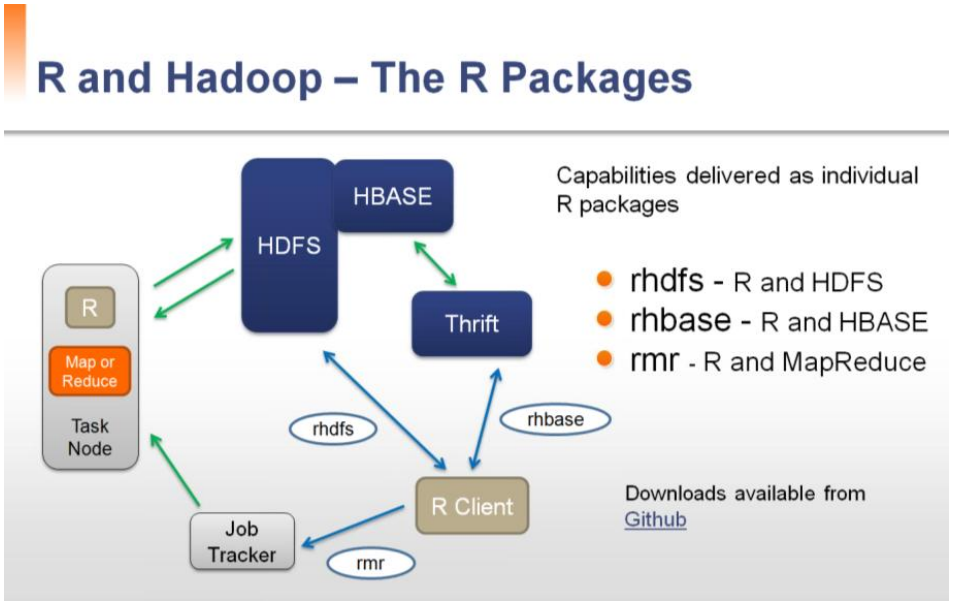
Analytics Techniques need to be written in Map-Reduce to execute on the clusters

Apache Mahout provides scalable machine learning libraries on the hadoop platform, which can be used for large datasets.

Mahout currently has implementations for the following techniques:

- ✓Collaborative Filtering
- ✓User and Item based recommenders
- ✓K-Means, Fuzzy K-Means clustering
- ✓Mean Shift clustering
- ✓Dirichlet process clustering
- ✓Latent Dirichlet Allocation
- ✓Singular value decomposition
- ✓Parallel Frequent Pattern mining
- ✓Complementary Naive Bayes classifier
- ✓Random forest decision tree based classifier
- ✓High performance java collections (previously colt collections)

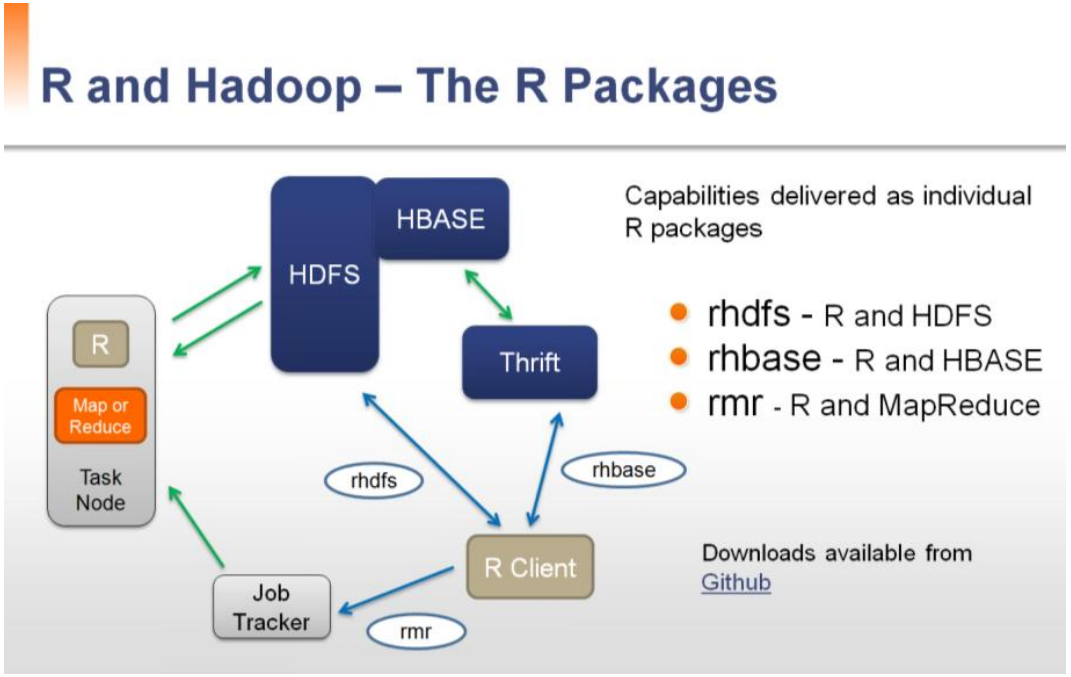
✓http://mahout.apache.org/

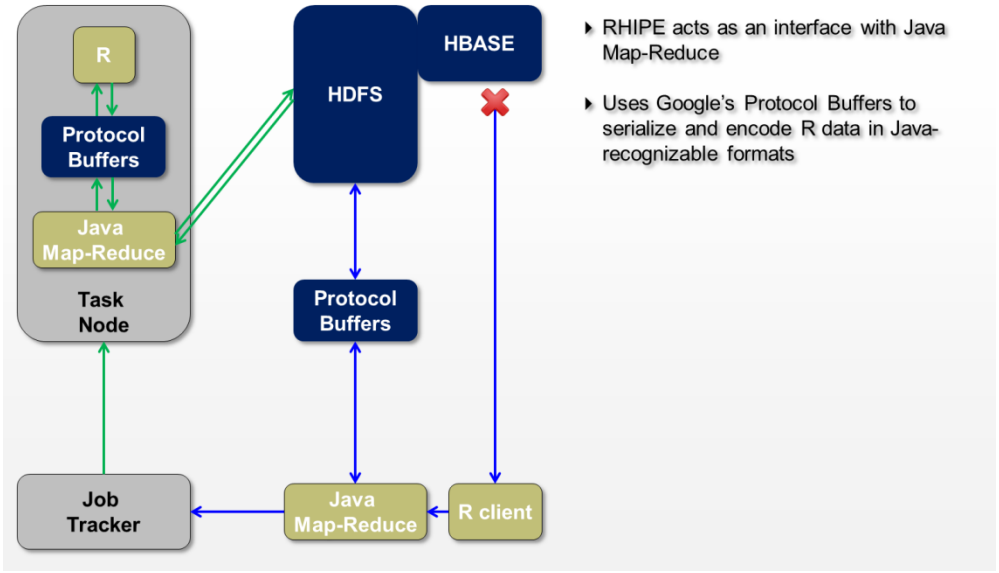


- ✓Recently released by Revolution Analytics
- ✓Designed for R programmers
- ✓Allows for Map Reduce jobs using the R language

R Packages for Map-Reduce

- Open source APIs for the Hadoop framework allowing users to define and run map-reduce jobs in R
- **RHADOOP**
  - Recent release by Revolution analytics
  - Designed for R programmers
  - Allows for more intuitive map-reduce programming
- **RHIPE**
  - Frequently updated with an active community
  - Uses Google's Protocol buffers to serialize data
  - Based on Hadoop streaming source





Agenda

- What is Big Data?
- Introduction to the Hadoop ecosystem
- Analytics using Hadoop, R, Mahout
- Competency requirements by focus area

Competency requirements by focus area

Focus Area: Data Extraction	Focus: Data Transformation	Focus Area: Modeling
Pig	Pig	R
Hive	Hive	Mahout library
RHive	RHive	Java
Java	Java	Understanding of Map-Reduce
Understanding of Map-Reduce	R	
	Understanding of Map-Reduce	

Appendix

A quick glossary of Hadoop terminologies

Hadoop	**Master unit**	Controlling unit in a cluster Assigns map jobs to task units
	Task unit	Workhorses of the cluster Performs assigned map job on parts of data stored on it
Database	**Hbase**	Distributed database designed to run on Hadoop
Maintenance	**Zookeeper**	Centralized configuration service and naming registry; keeps a track of what data's where
Monitoring	**Nagios**	Monitoring tool; useful for monitoring individual services/processes running on various nodes
	Ganglia	Another monitoring tool; very graphical and useful for cluster-level monitoring
Analytics	**R**	Statistical programming language
	RHIPE / RHadoop	APIs for Hadoop, allowing users to define and run map-reduce jobs in R
	Protocol Buffers (for RHIPE)	Serialization format by google; enables RHIPE to interpret hadoop data as R objects
	Pig/Hive	Programming language specifically for Map-Reduce
Performance	**Fair scheduler**	A scheduler for hadoop; enables optimal allocation of computation resources

Useful links

- http://hadoop.apache.org
- http://mahout.apache.org/
- http://wiki.apache.org/hadoop
- http://en.wikipedia.org/wiki/Apache_Mahout
- http://wiki.apache.org/hadoop/HadoopIsNot
- http://pig.apache.org/
- http://hadoop.apache.org/hdfs/

Thank You

Chicago, IL
Bangalore, India
March 15, 2012
www.mu-sigma.com