Big Data
Hadoop Stack

Lecture #1

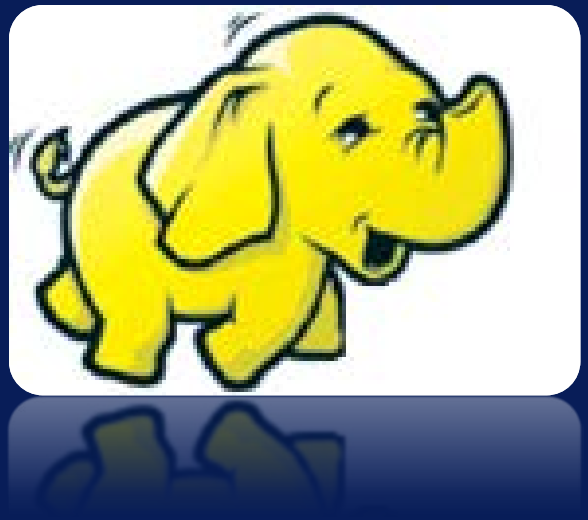
Hadoop Beginnings

What is Hadoop?

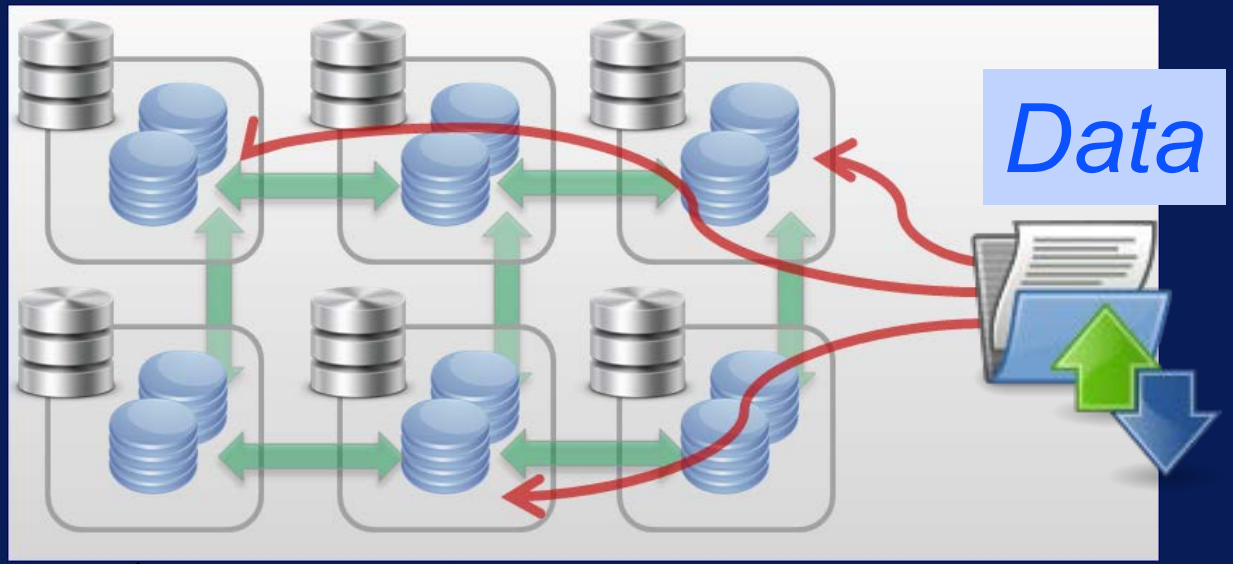
Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware

Hadoop was created by Doug Cutting and Mike Cafarella in 2005

Named the project after son's toy elephant



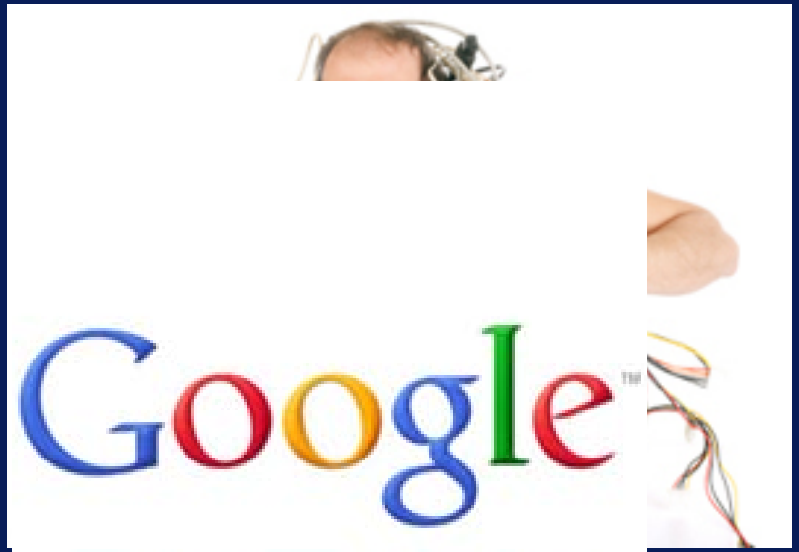
Moving Computation to Data



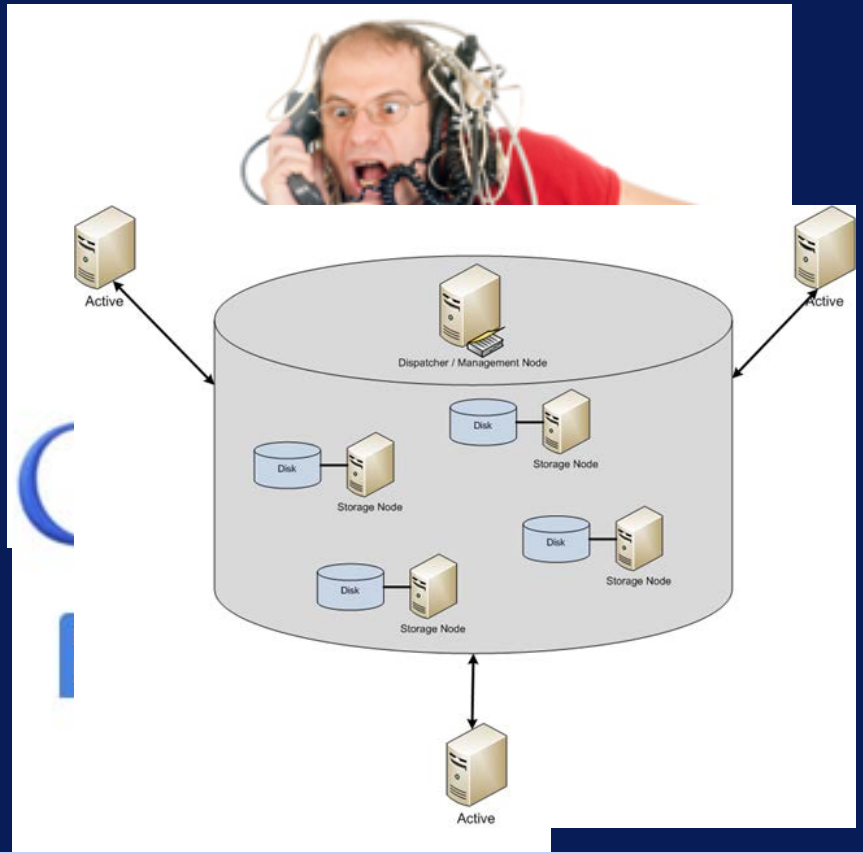
Computation

Scalability at Hadoop's core!

Reliability!
Reliability!
Reliability!



Reliability!
Reliability!
Reliability!



Reliability!
Reliability!
Reliability!

Google File System

Once
a year

365 Computers

Once
a day

Hourly

New Approach to Data

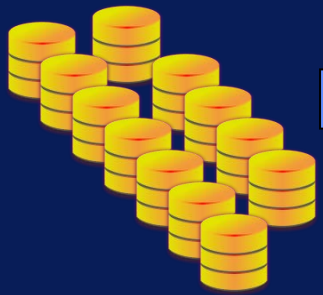
Keep all data

New Kinds of Analysis

Schema-on read style

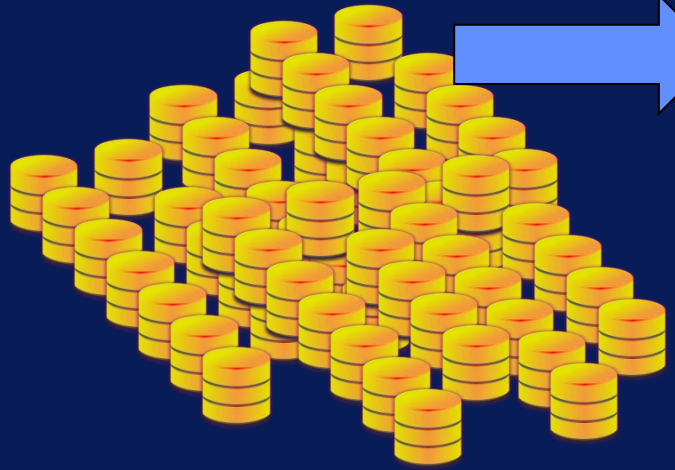
New Kinds of Analysis

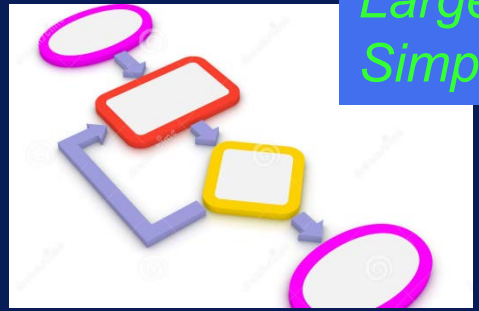
- *New analysis*



Small Data & Complex Algorithm

Vs.





Large Data & Simple Algorithm

Lecture #2

Apache Framework Hadoop Modules

Apache Framework Basic Modules

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop YARN
- Hadoop MapReduce

Apache Framework Basic Modules

Hadoop Common

Hadoop Distributed File System (HDFS)

Hadoop YARN

Hadoop MapReduce

Apache Framework Basic Modules

Hadoop Common

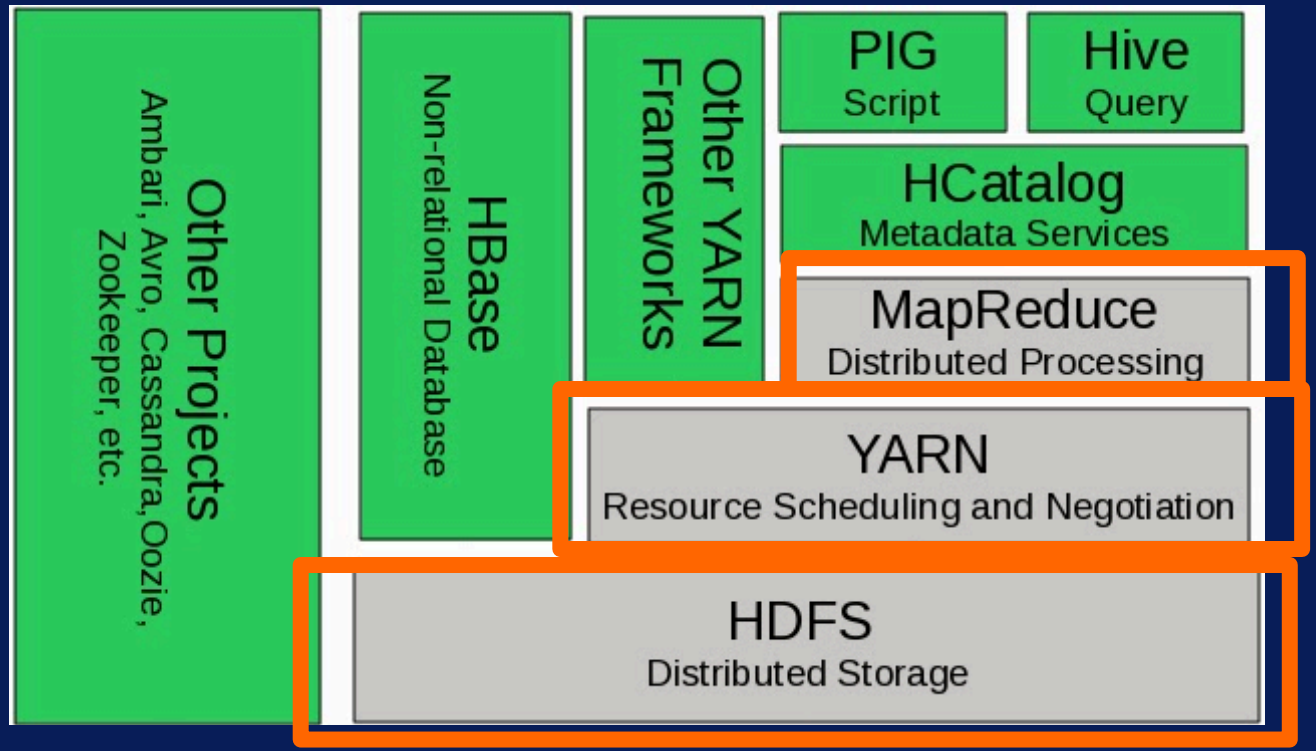
Hadoop Distributed File System (HDFS)

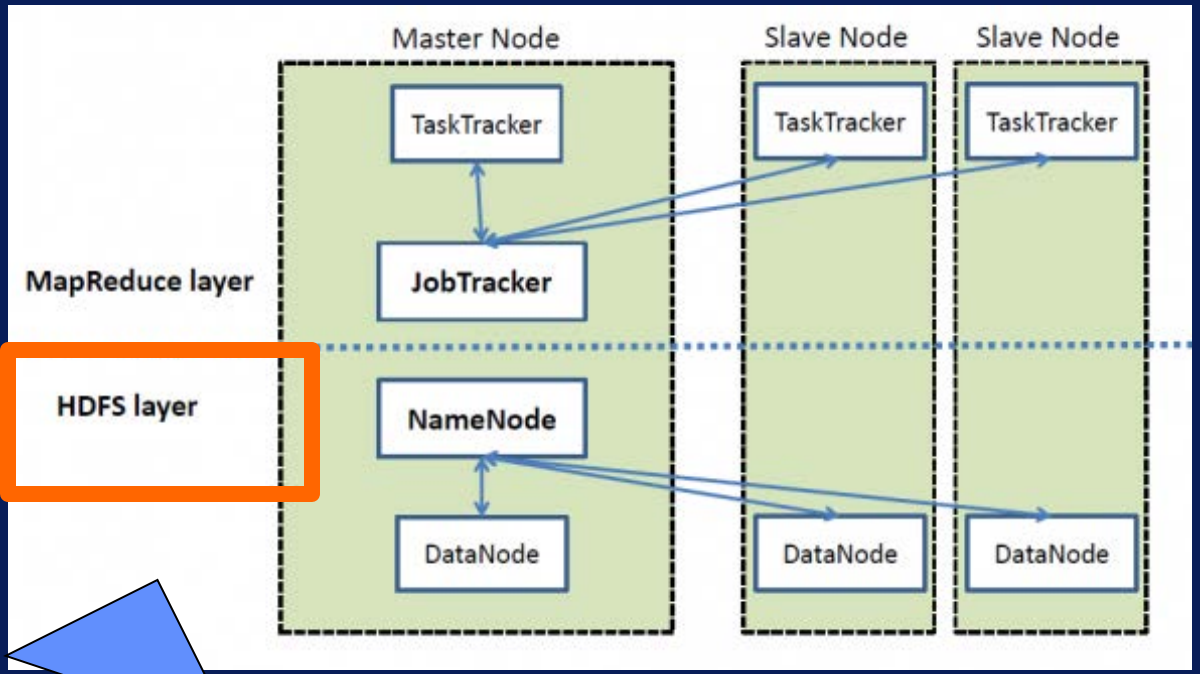
Hadoop YARN

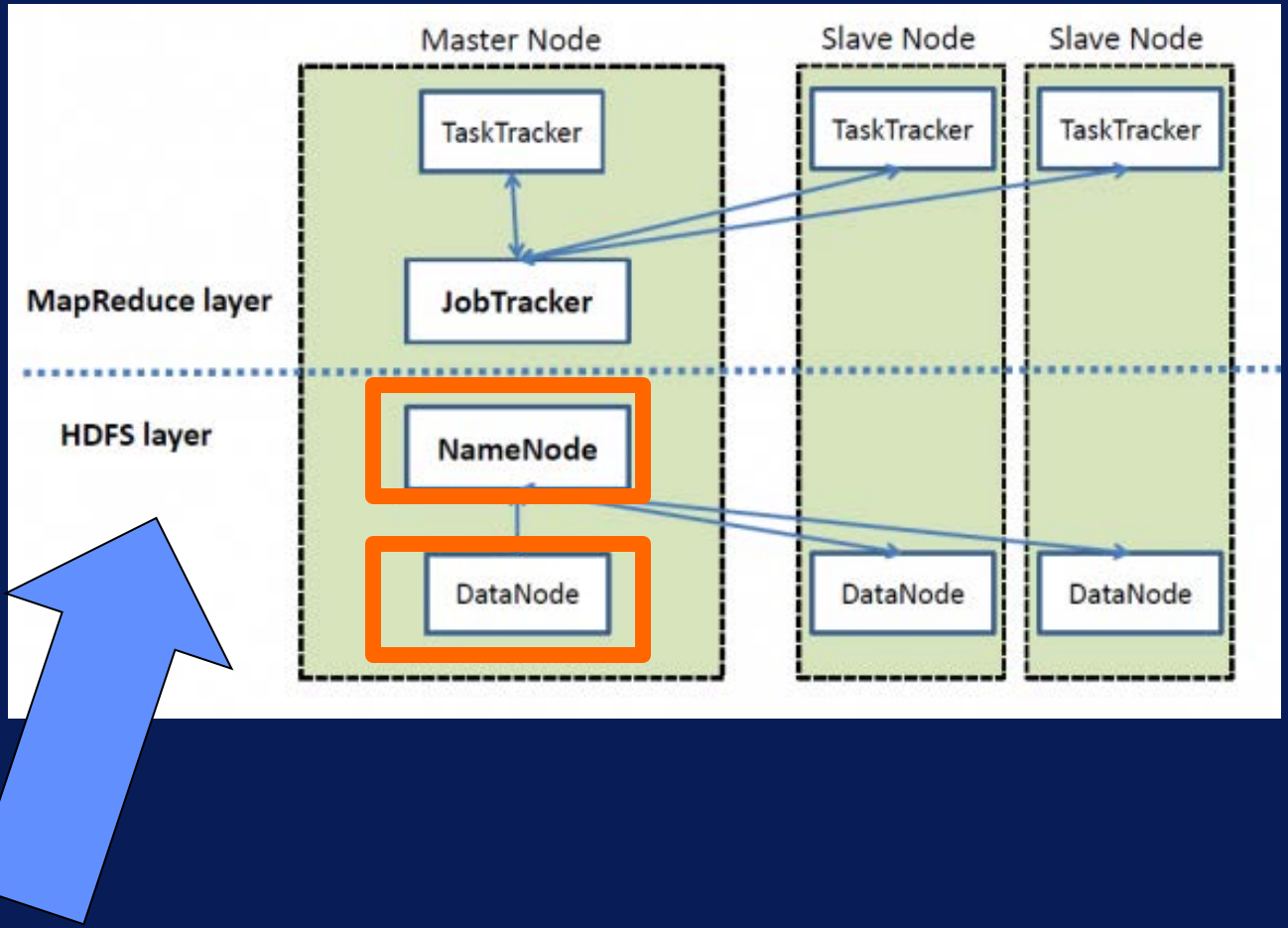
Hadoop MapReduce

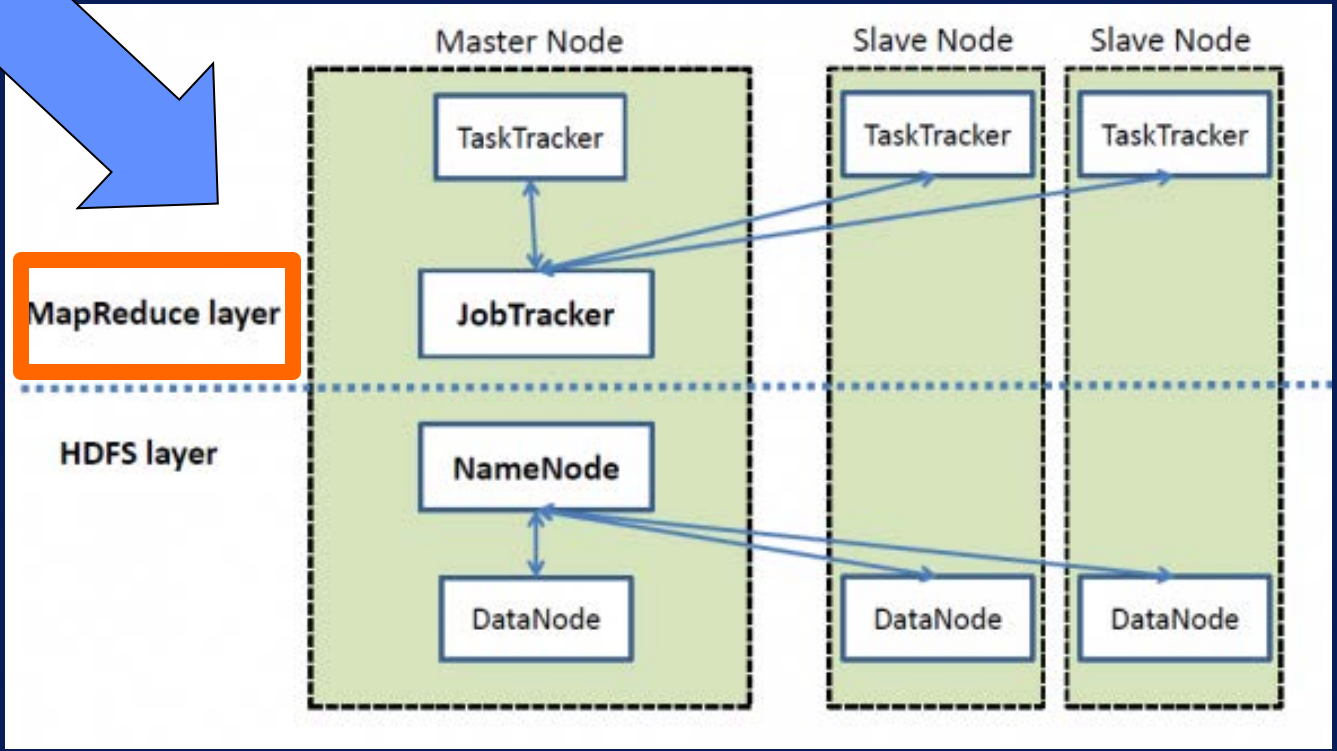
Apache Framework Basic Modules

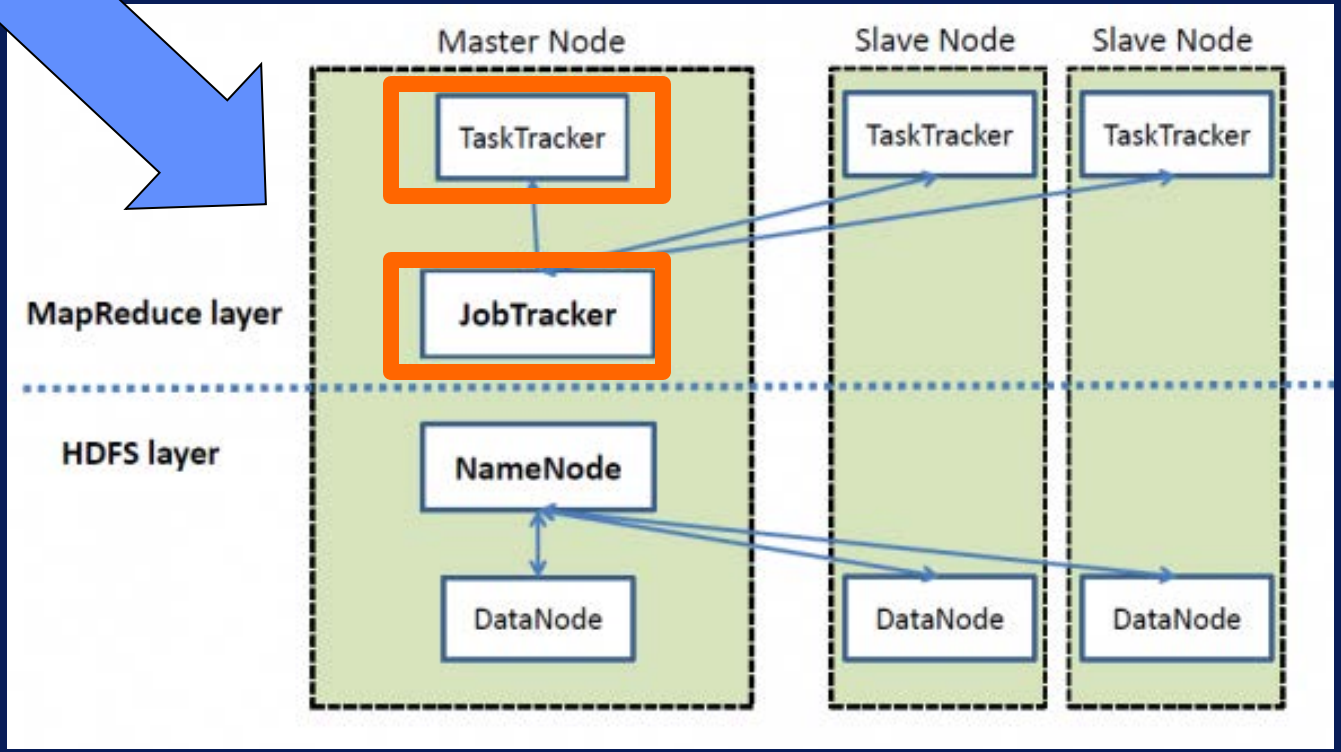
- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop YARN
- Hadoop MapReduce


Other Projects
Ambari, Avro, Cassandra, Oozie, Zookeeper, etc.
HBase
Non-relational Database
Other YARN Frameworks
PIG
Script
Hive
Query
HCatalog
Metadata Services
MapReduce
Distributed Processing
YARN
Resource Scheduling and Negotiation
HDFS
Distributed Storage


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
MapReduce layer
JobTracker
HDFS layer
NameNode
DataNode
DataNode
DataNode


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
JobTracker
MapReduce layer
HDFS layer
NameNode
DataNode
DataNode
DataNode


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
MapReduce layer
JobTracker
HDFS layer
NameNode
DataNode
DataNode
DataNode


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
JobTracker
MapReduce layer
HDFS layer
NameNode
DataNode
DataNode
DataNode

Lecture #3

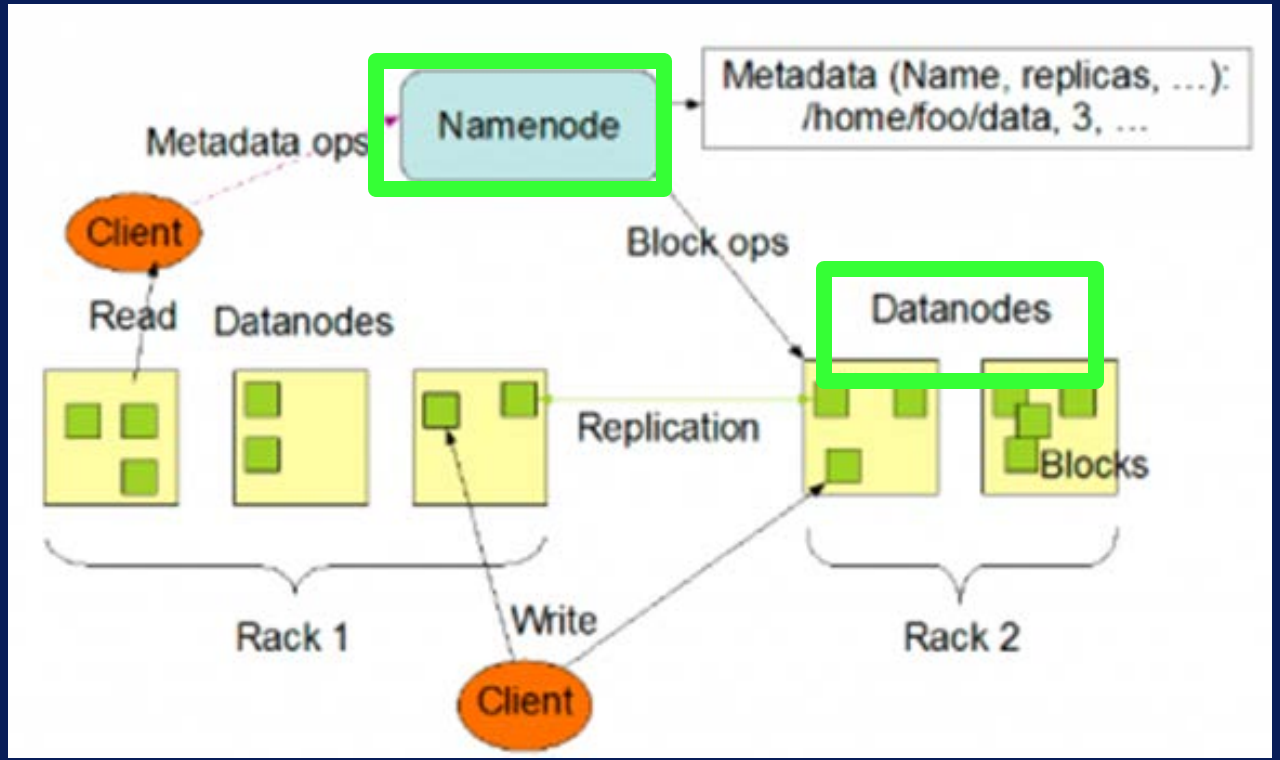
Hadoop Distributed File System (HDFS)

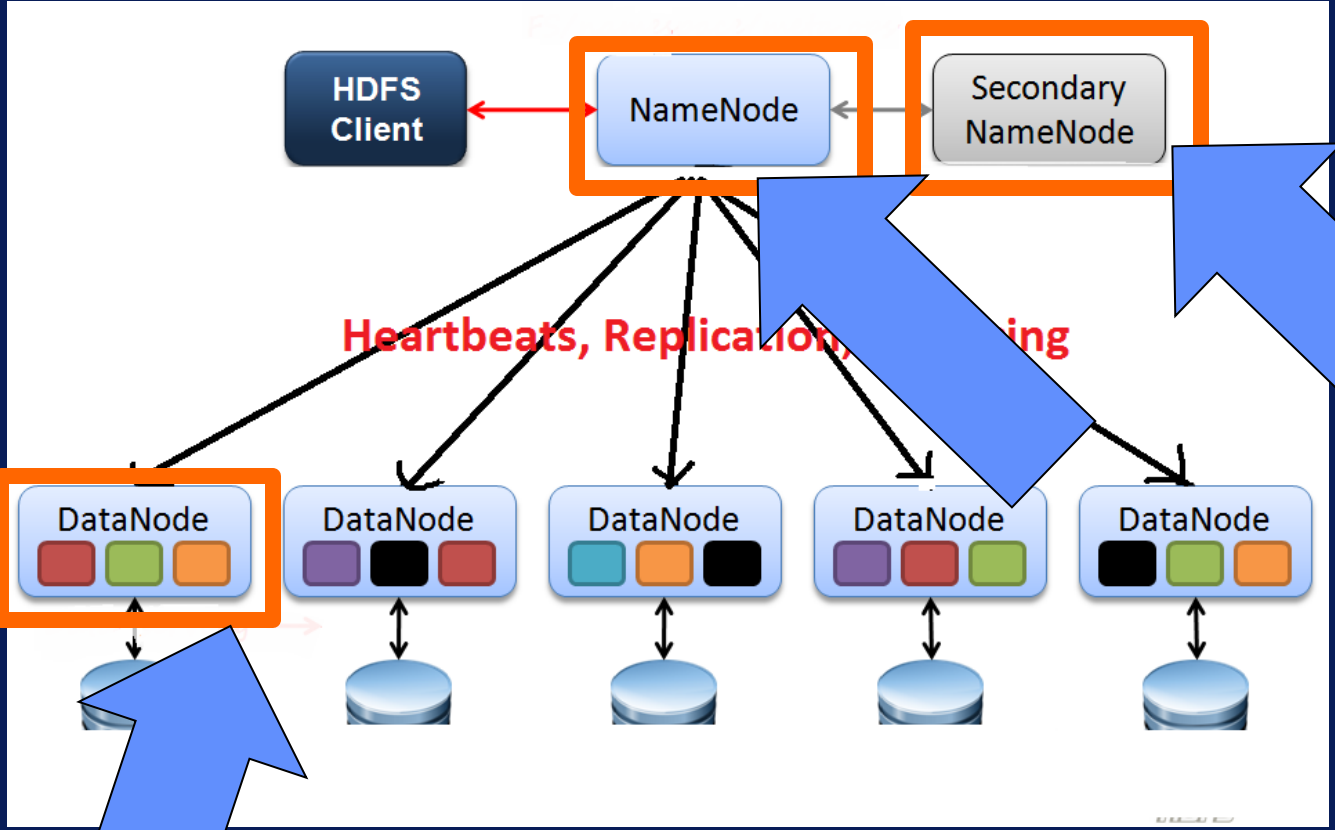
HDFS

Hadoop Distributed File System

Distributed, scalable, and portable file-system written in Java for the Hadoop framework

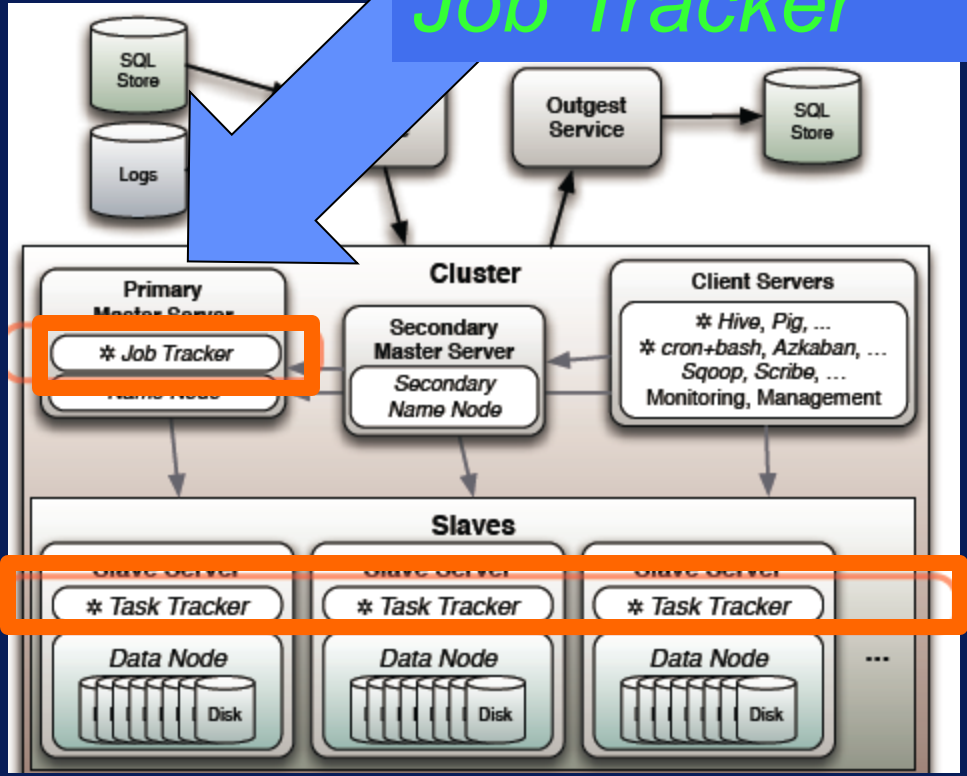
HDFS


Namenode
Metadata (Name, replicas, ...):
/home/foo/data, 3, ...
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

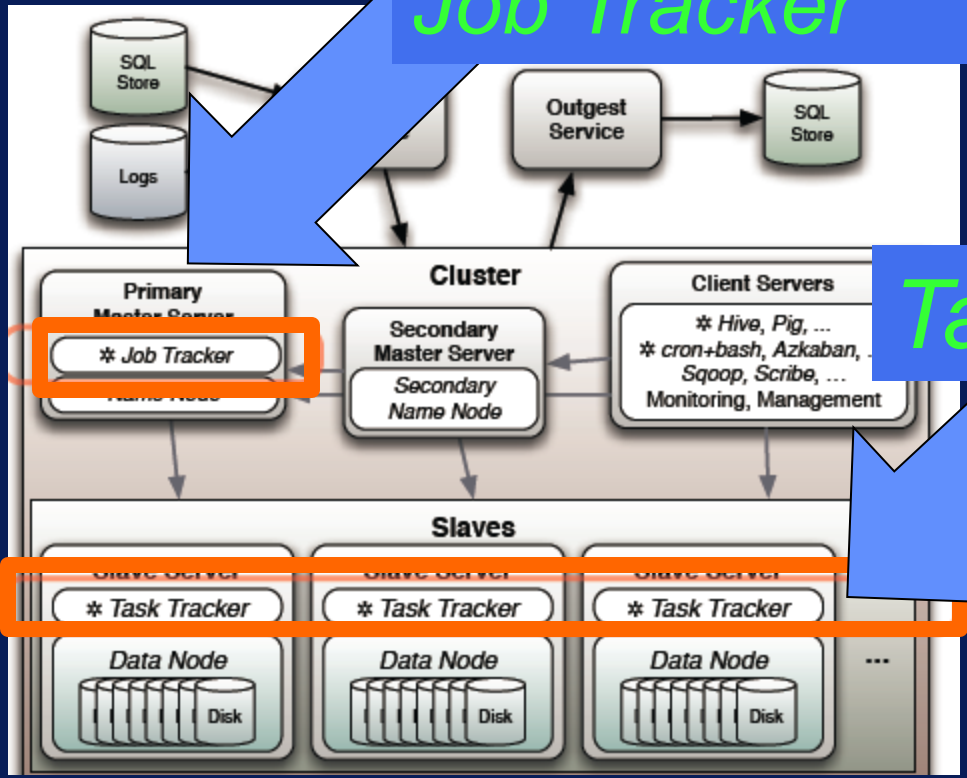

HDFS Client
NameNode
Secondary NameNode
Heartbeats, Replication,
ing
DataNode
DataNode
DataNode
DataNode
DataNode

MapReduce Engine

Job Tracker



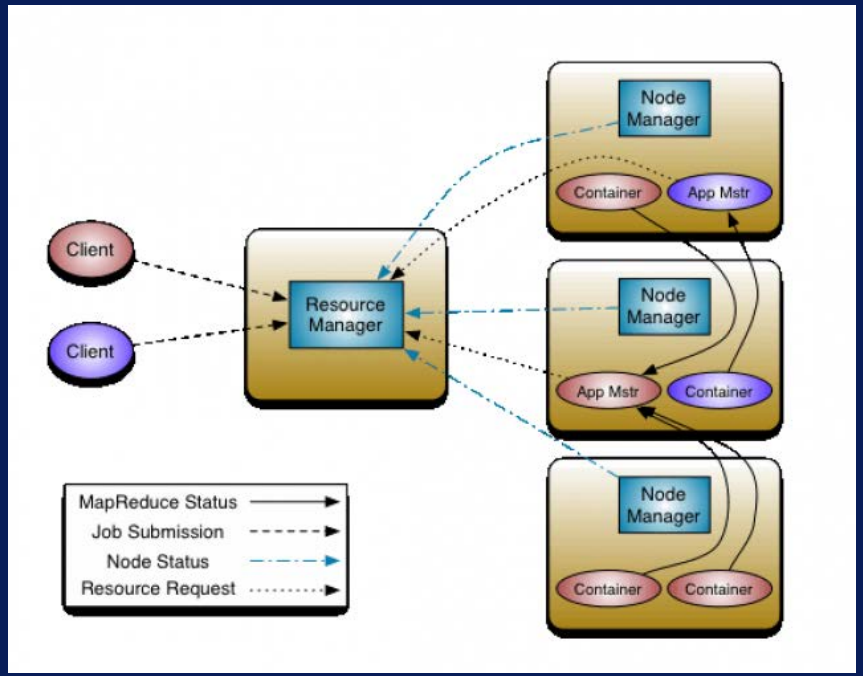
MapReduce Engine

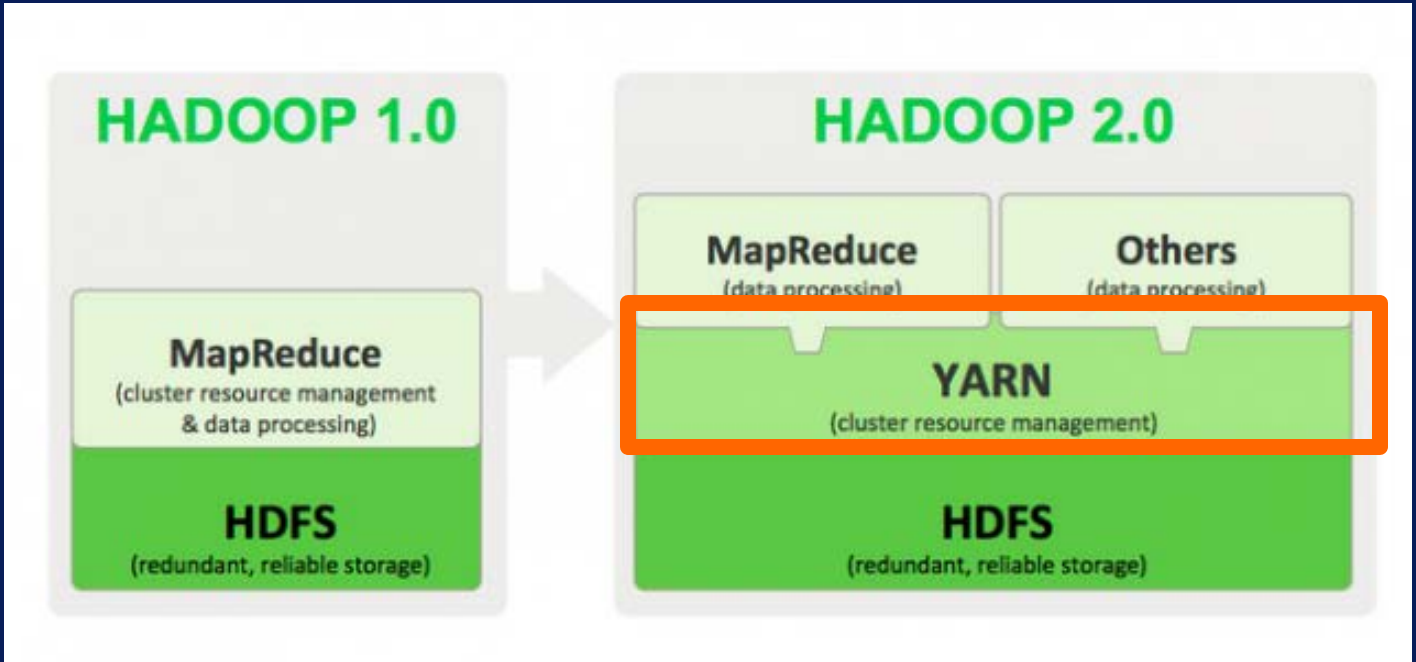


Job Tracker

Task Tracker

Apache Hadoop NextGen MapReduce (YARN)




HADOOP 1.0
MapReduce
(cluster resource management
& data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

What is Yarn?

- **YARN enhances the power of a Hadoop compute cluster**

Scalability

What is Yarn?

- **YARN enhances the power of a Hadoop compute cluster**

Scalability

Improved cluster utilization

What is Yarn?

- **YARN enhances the power of a Hadoop compute cluster**

Scalability

Improved cluster utilization

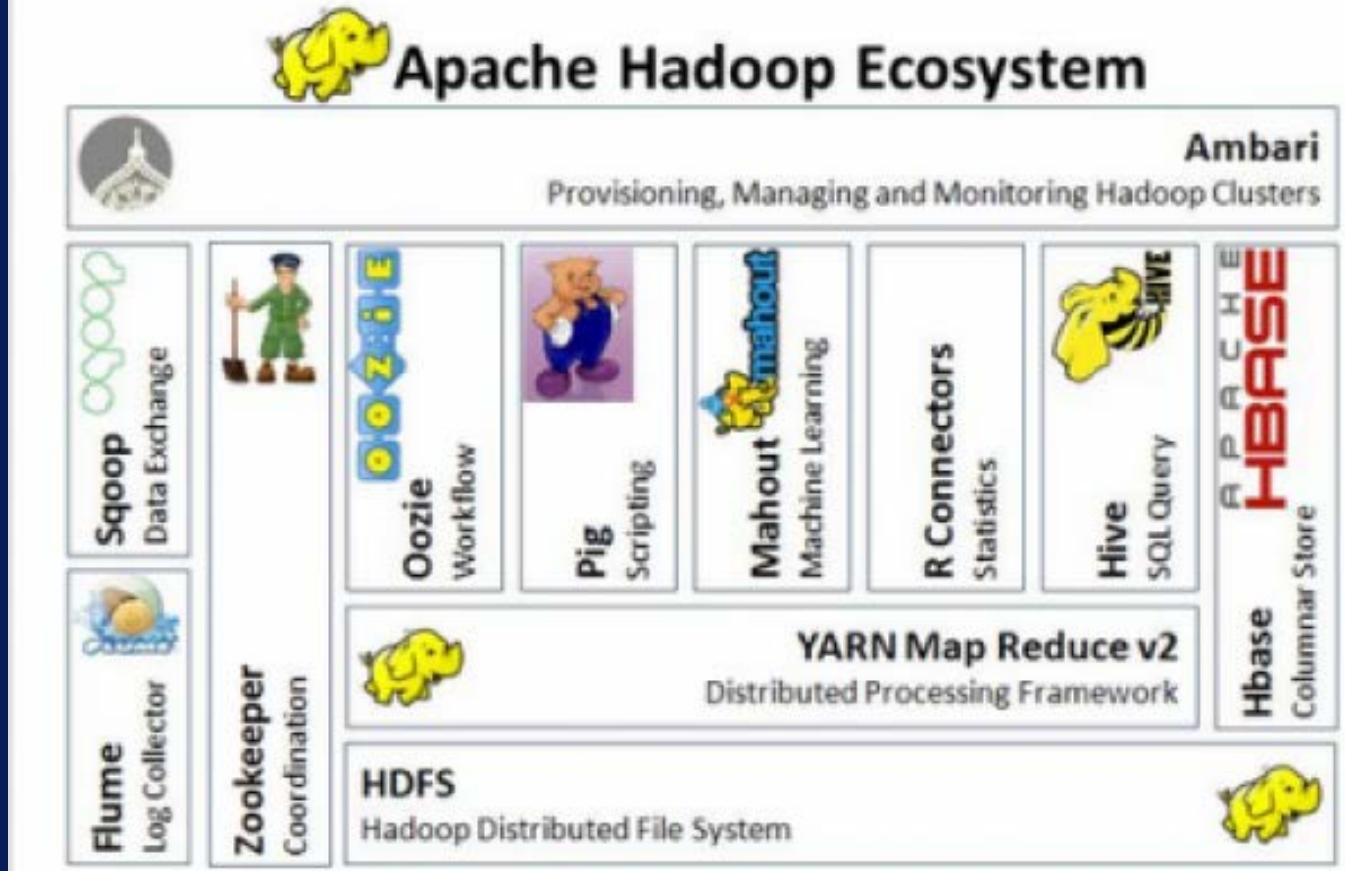
MapReduce Compatibility

What is Yarn?

- **YARN enhances the power of a Hadoop compute cluster**

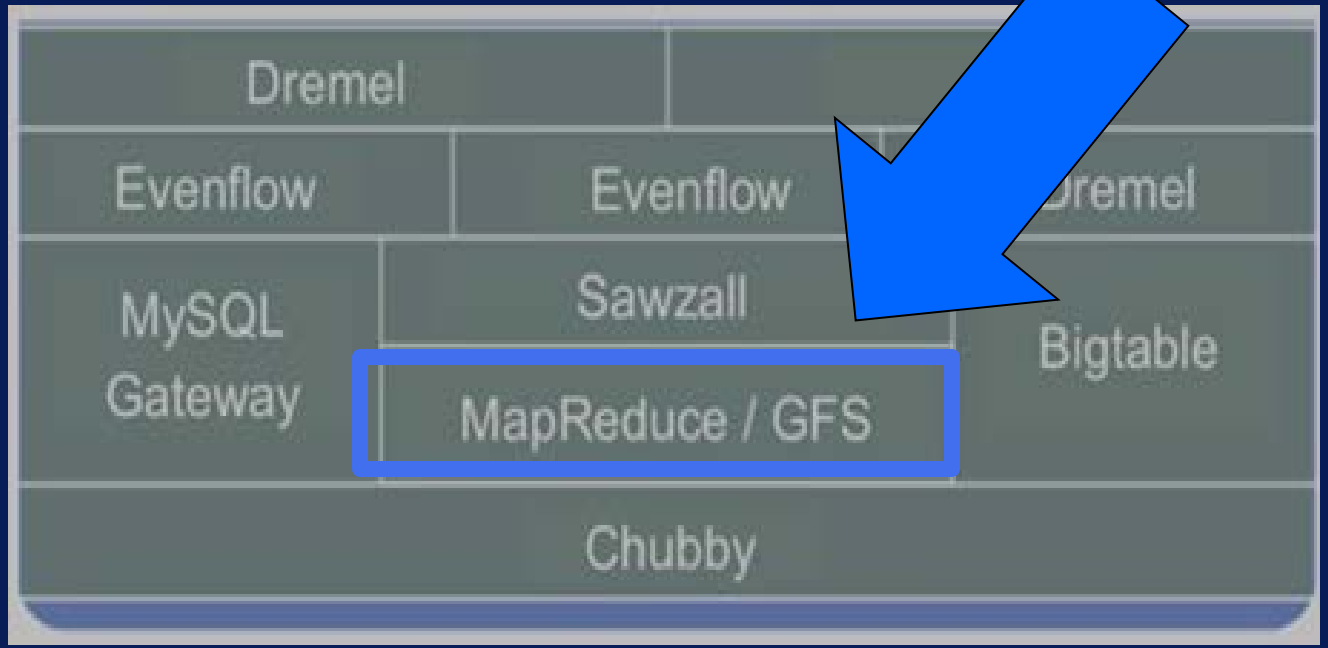
Lecture #4

The Hadoop “Zoo”


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

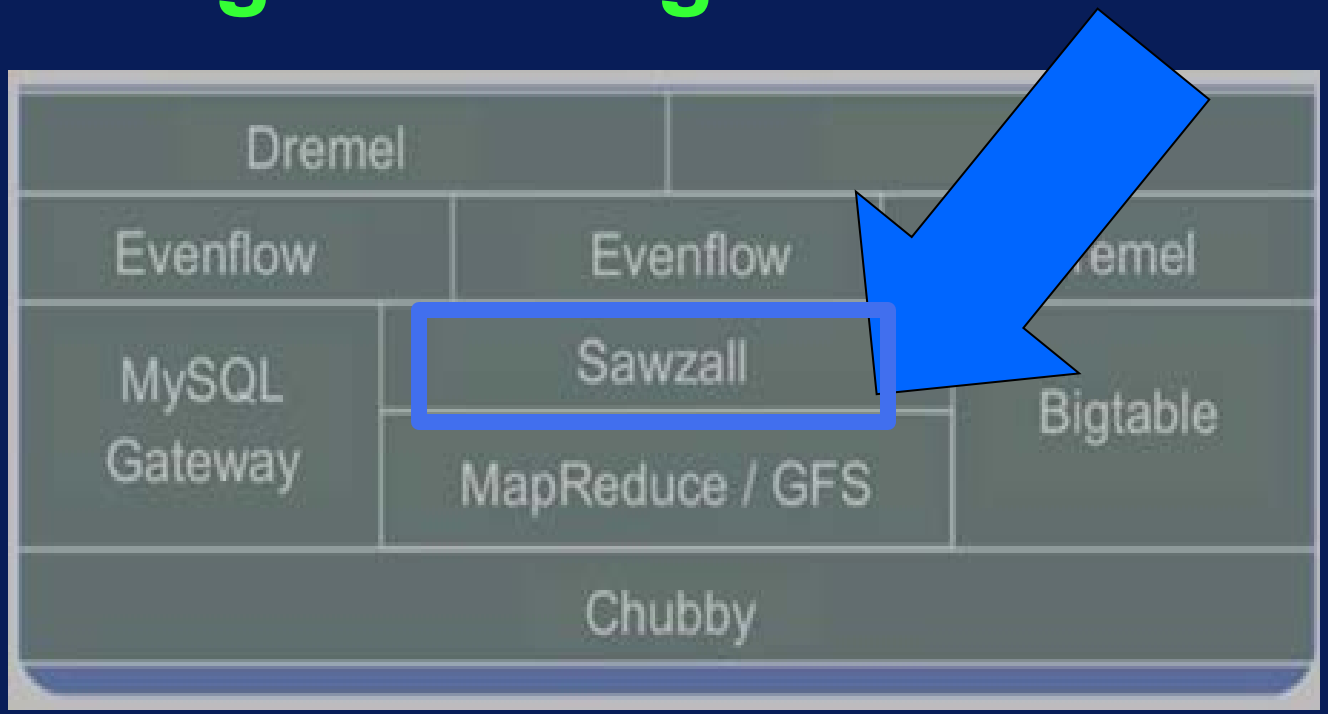
How to figure out the Zoo??

Original Google Stack



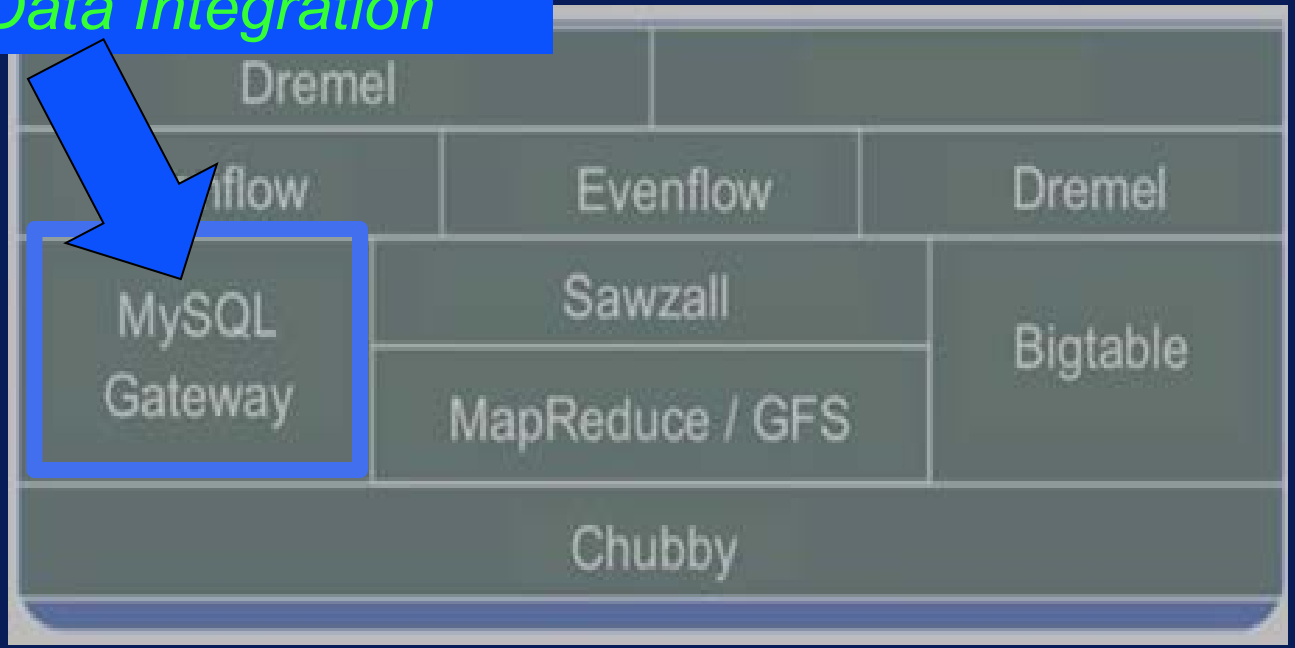
Original Google Stack

Original Google Stack

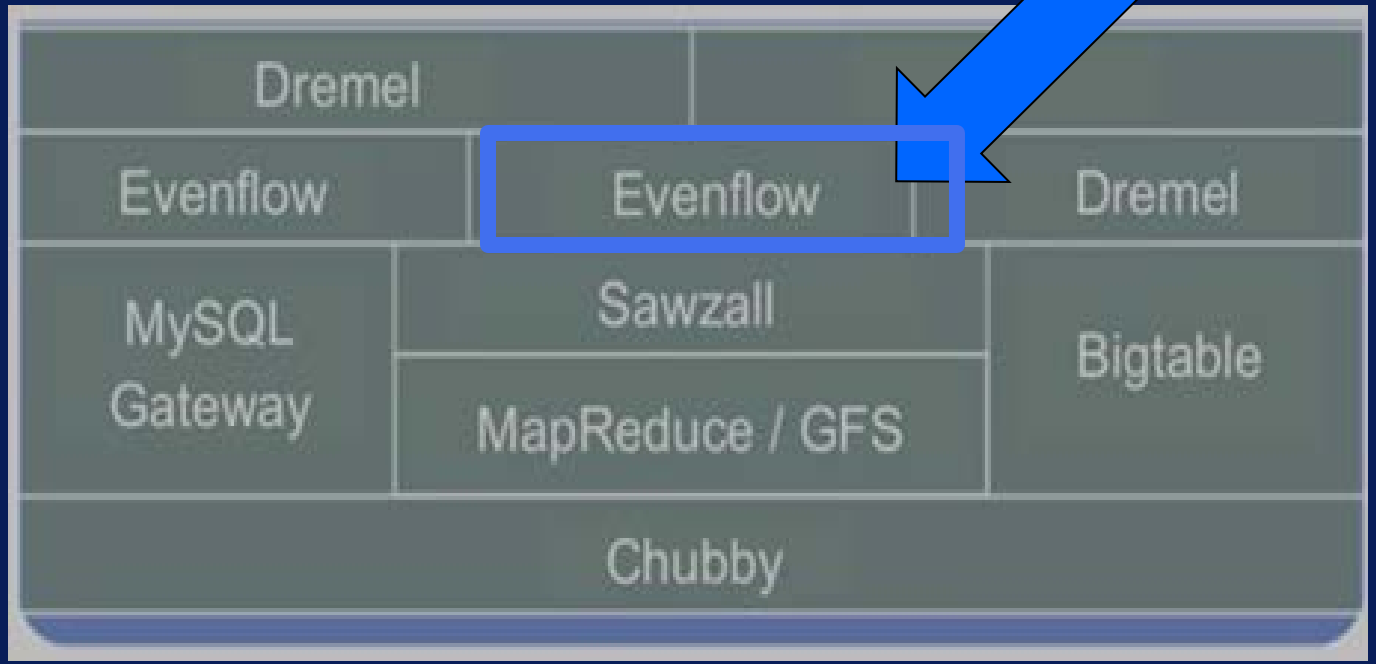


Original Google Stack

Data Integration



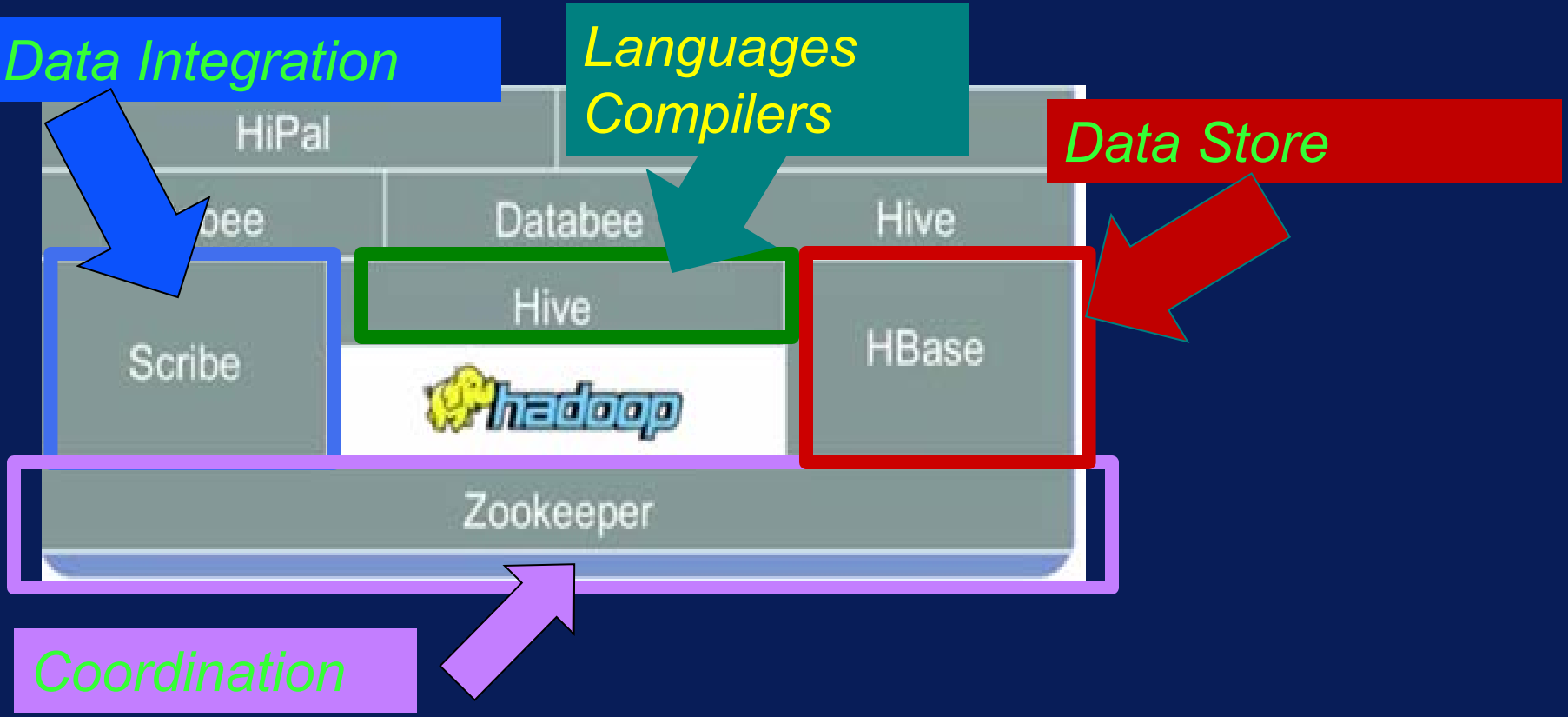
Original Google Stack



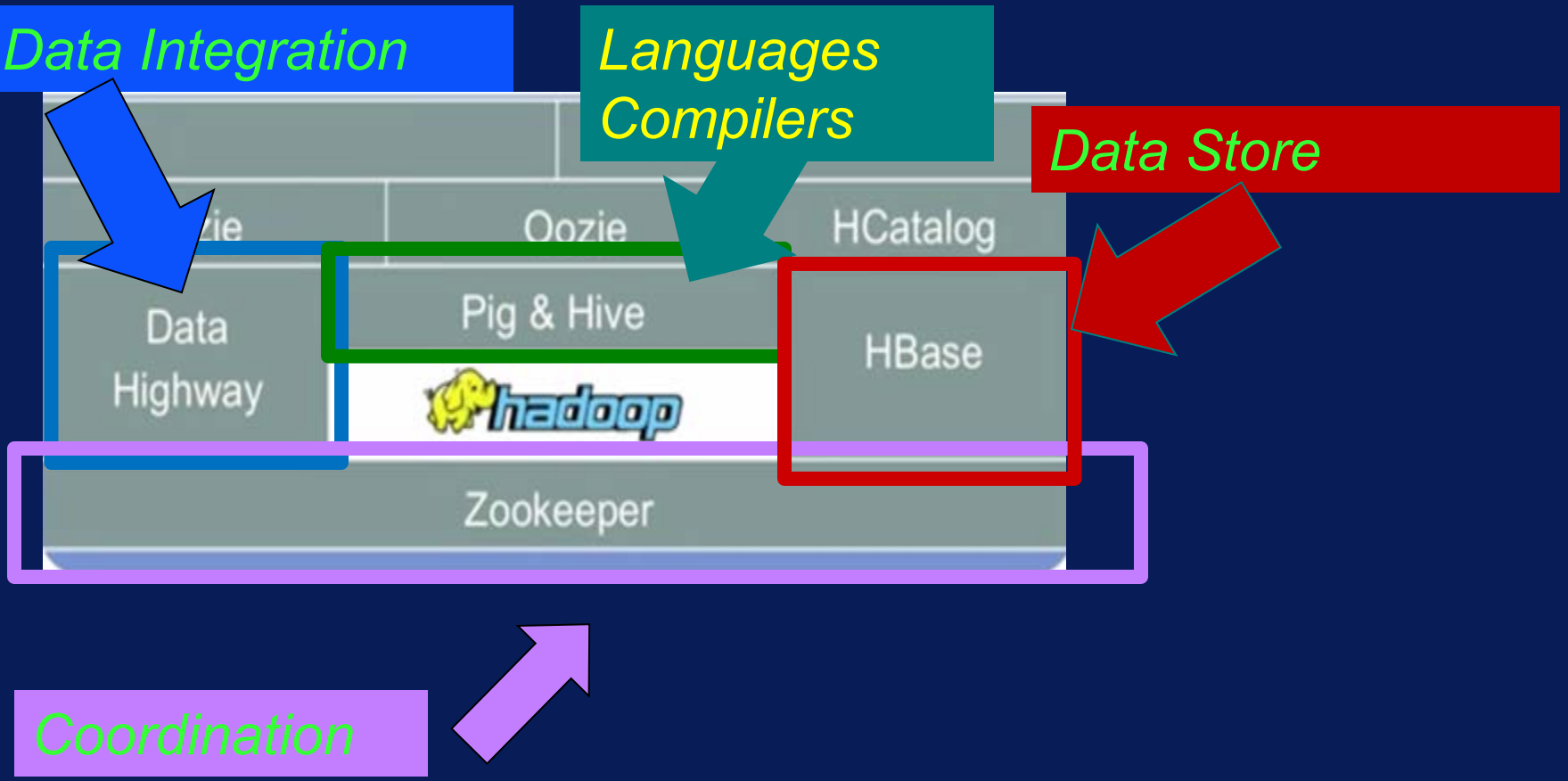
Original Google Stack

Original Google Stack

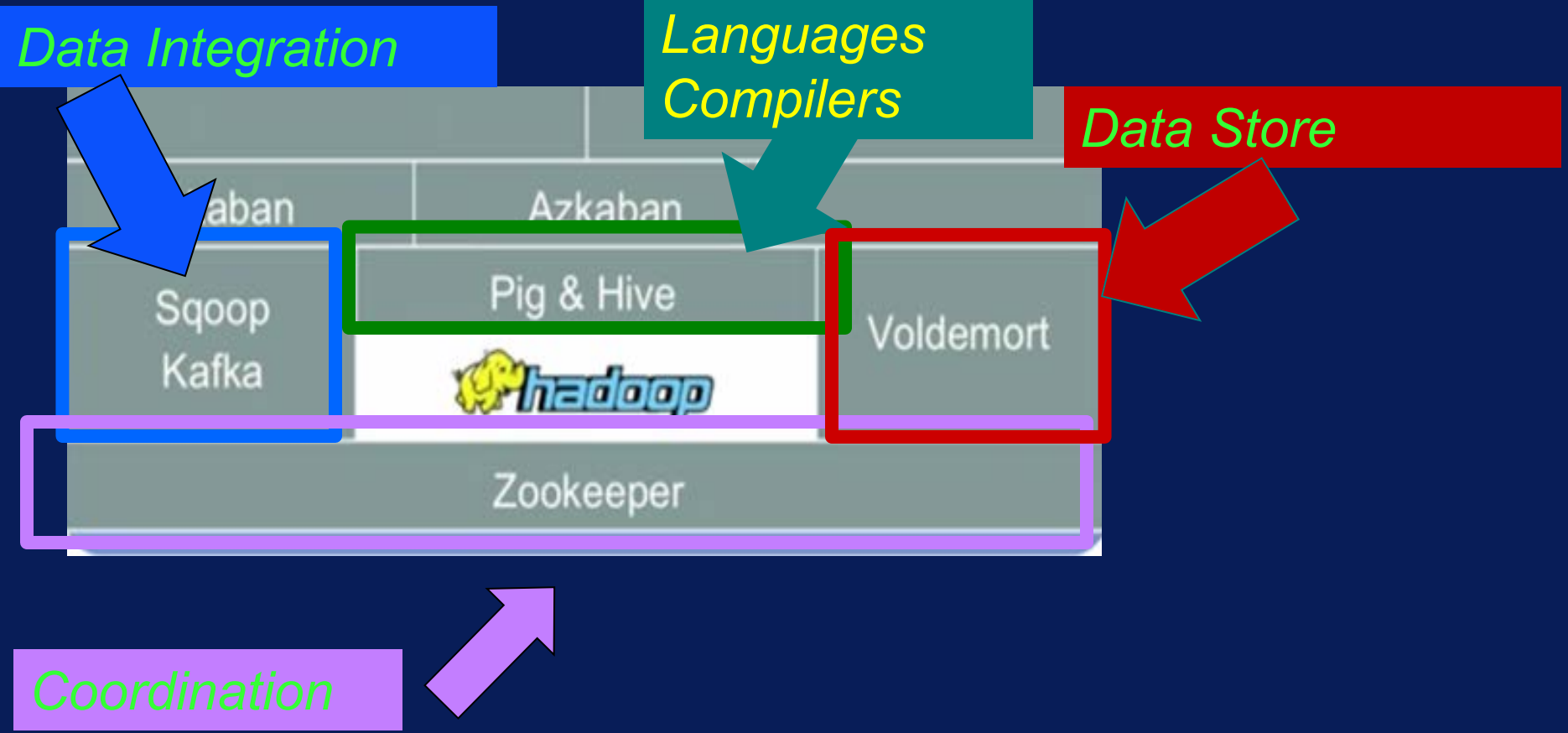
Facebook's Version of the Stack



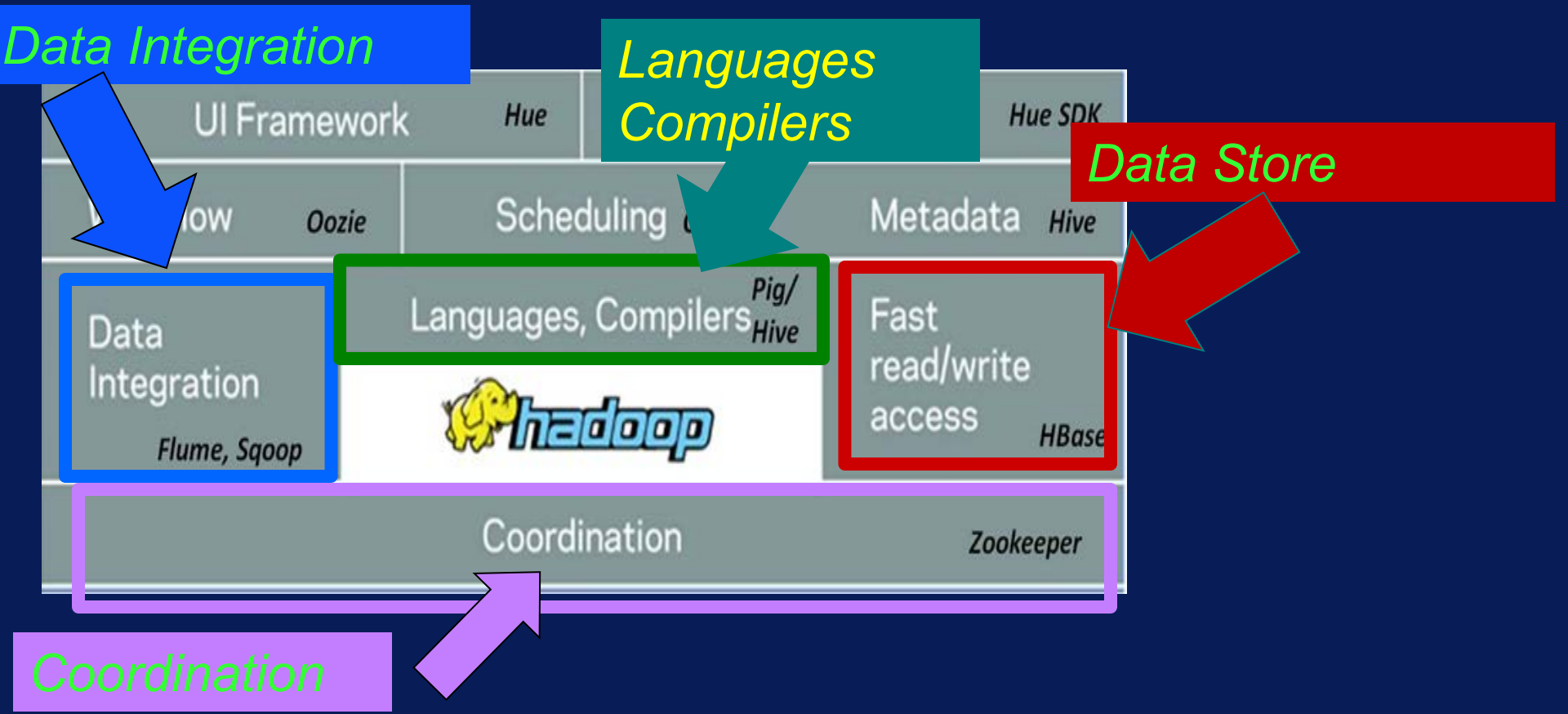
Yahoo's Version of the Stack



LinkedIn's Version of the Stack

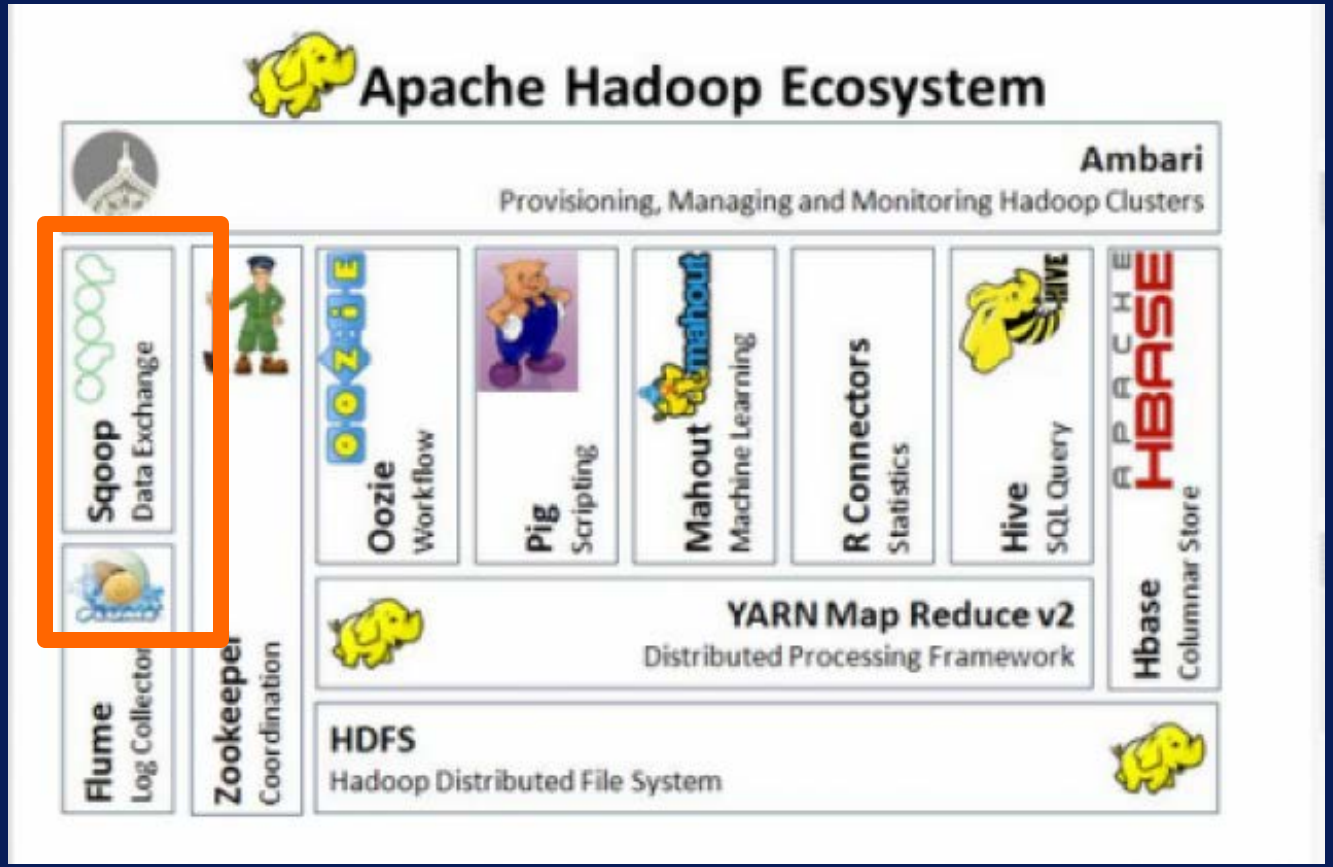


Cloudera's Version of the Stack



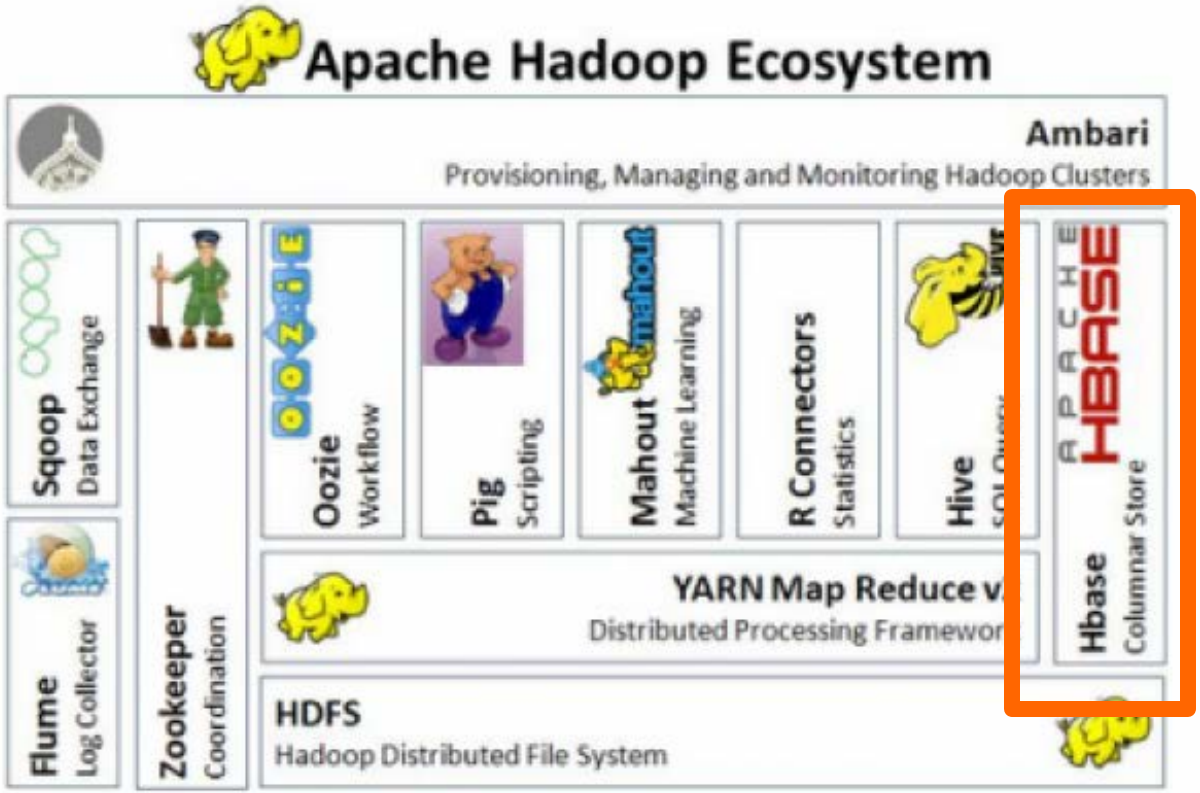
Lecture #5

Hadoop Ecosystem Major Components


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

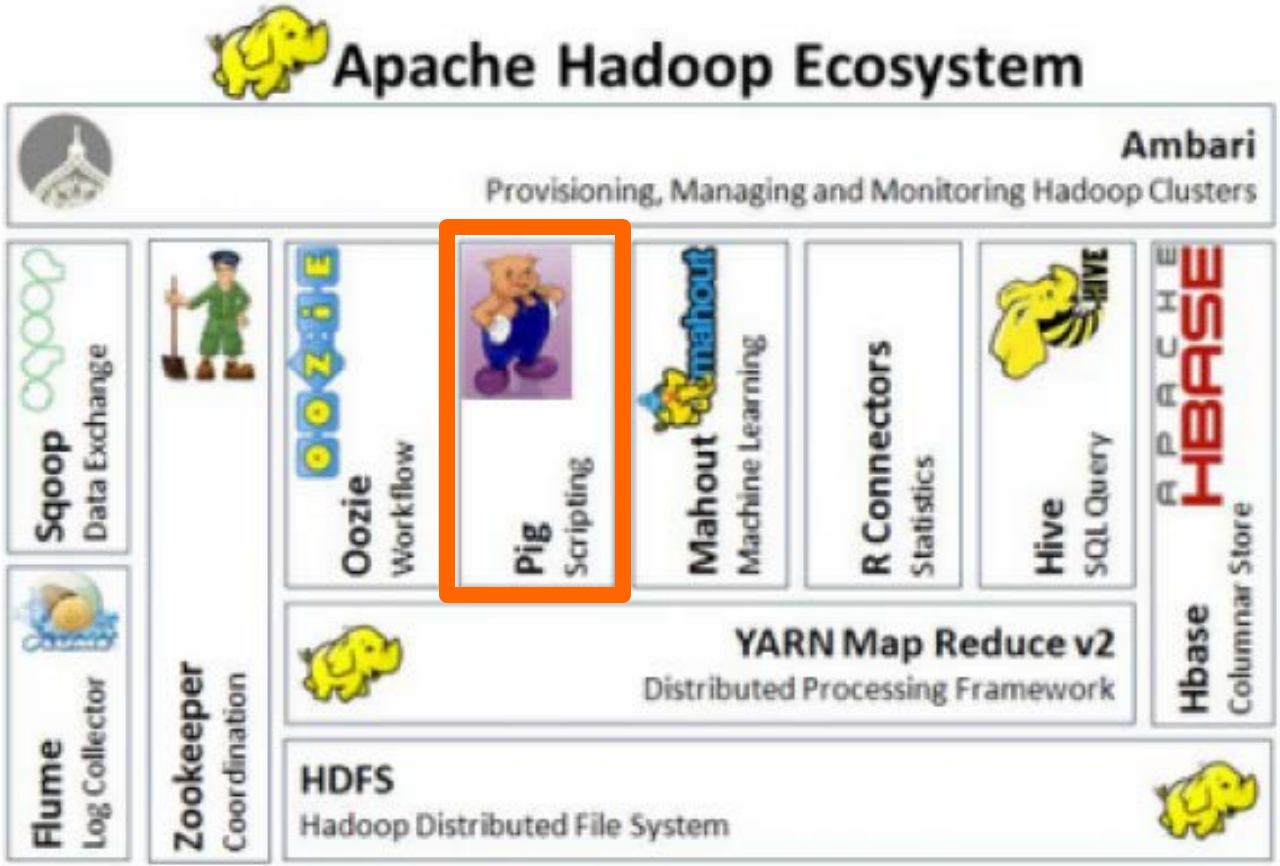
Apache Sqoop

- Tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases

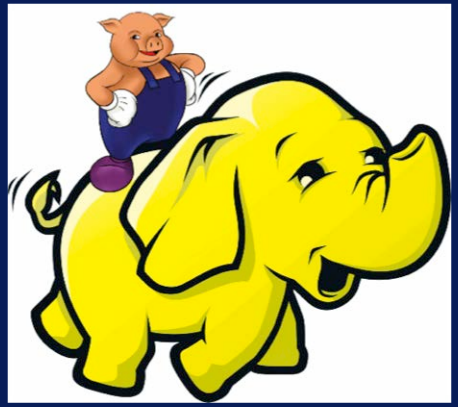

Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
Hbase
Columnar Store
YARN Map Reduce v
Distributed Processing Framewor
HDFS
Hadoop Distributed File System

HBASE

- Column-oriented database management system
- Key-value store
- Based on Google Big Table
- Can hold extremely large data
- Dynamic data model
- Not a Relational DBMS

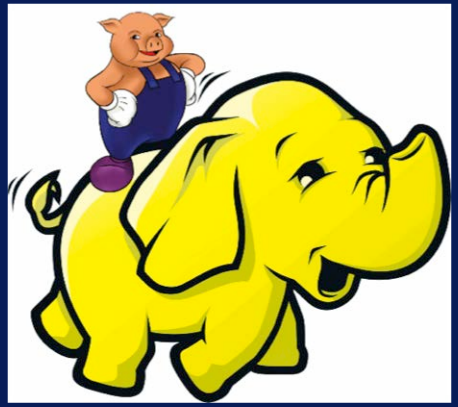

Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

PIG



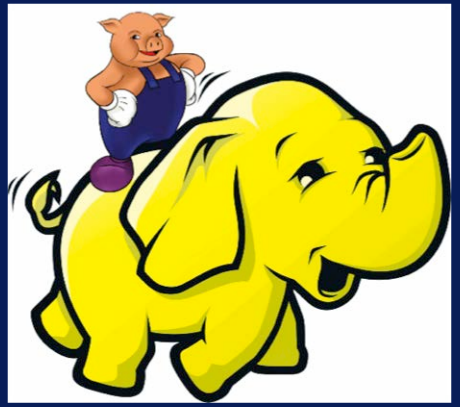
High level programming on top of Hadoop MapReduce

PIG



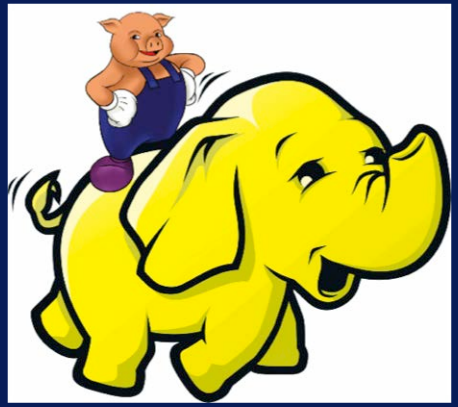
The language: Pig Latin

PIG



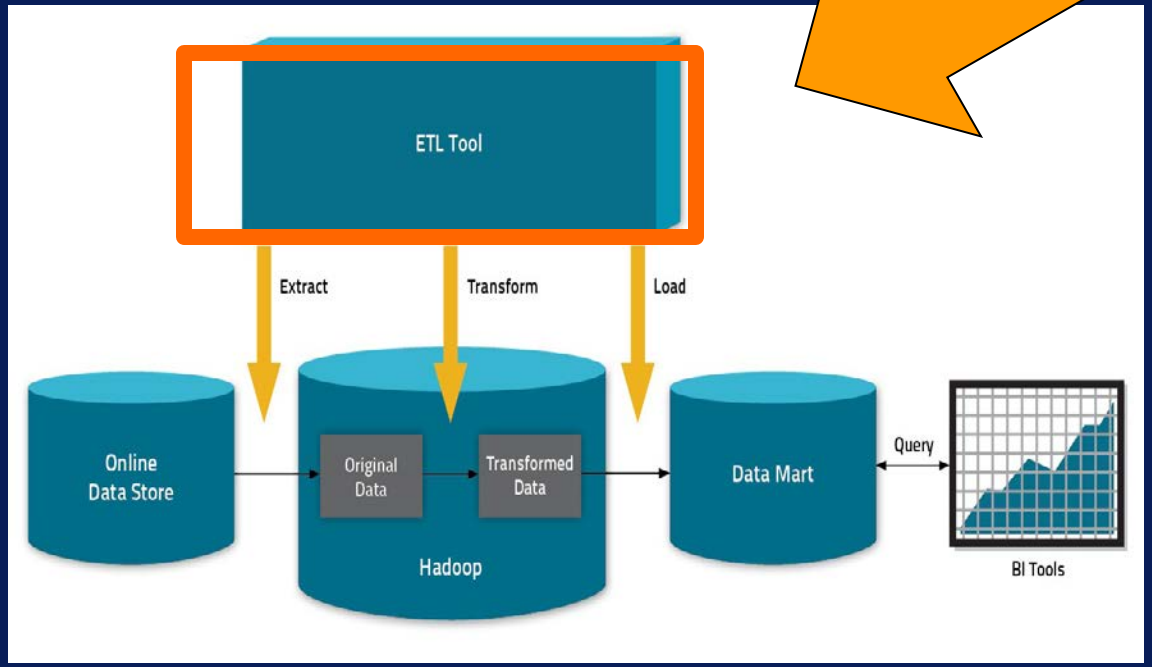
Data analysis problems as data flows

PIG

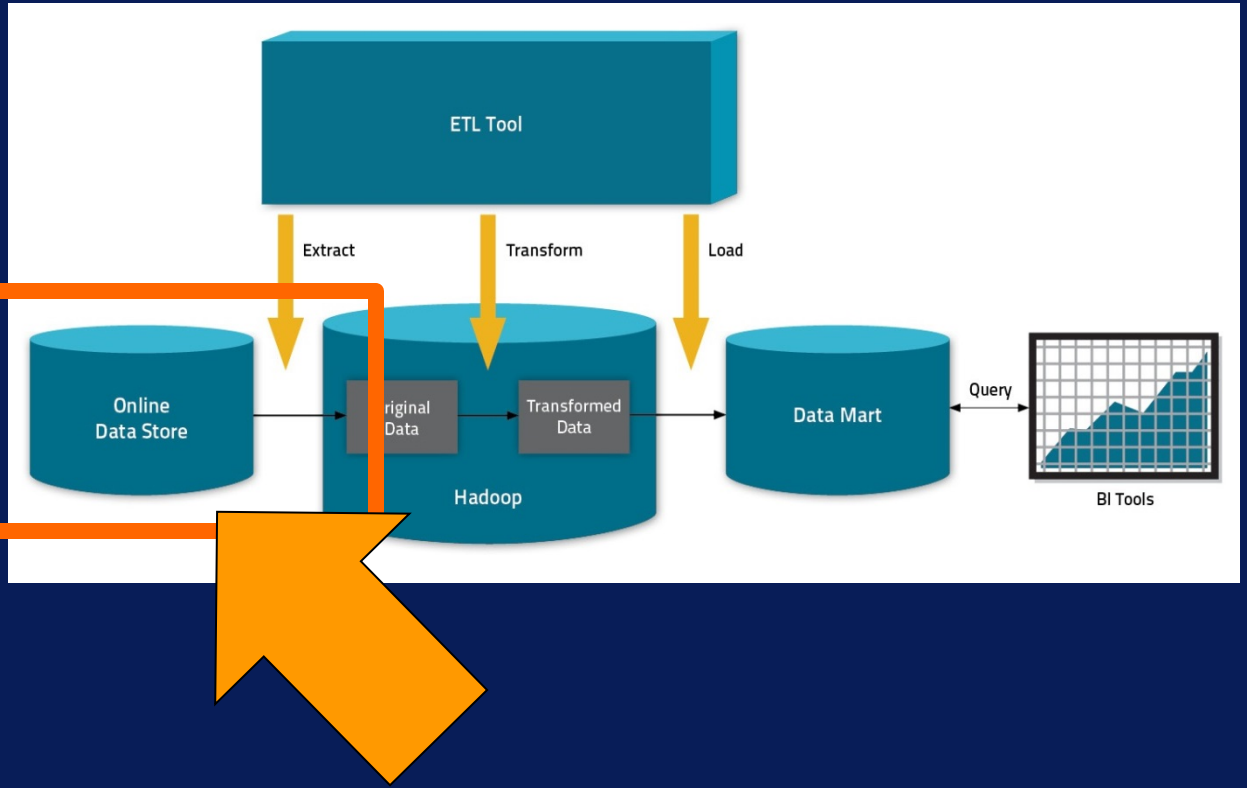


Originally developed at Yahoo 2006

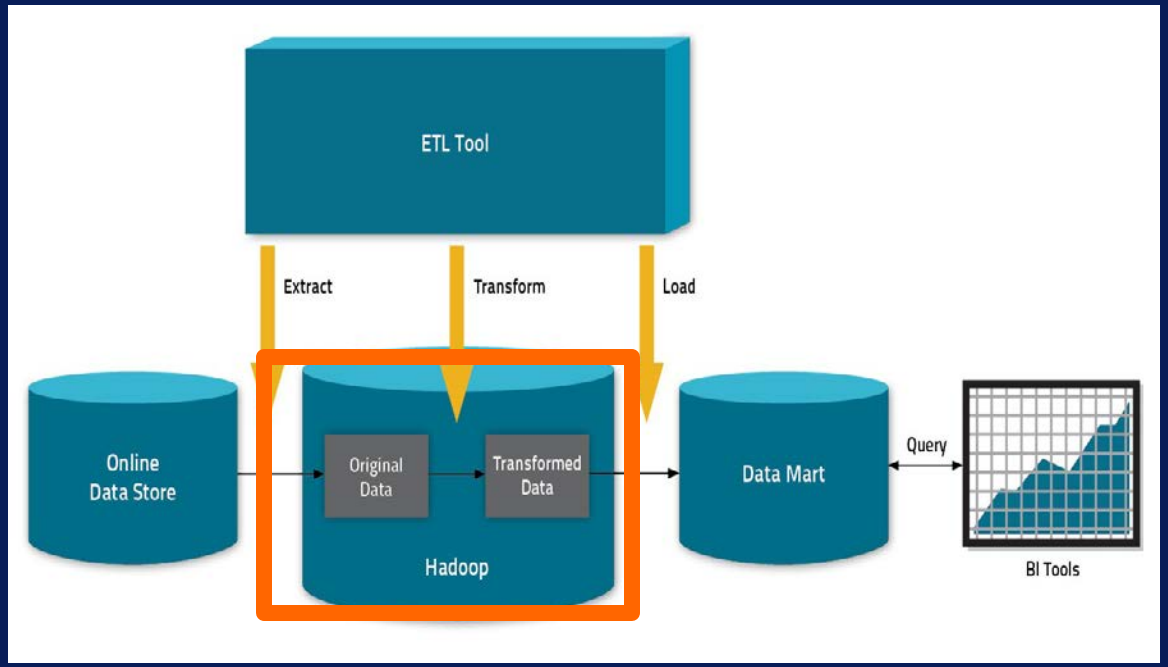
Pig for ETL

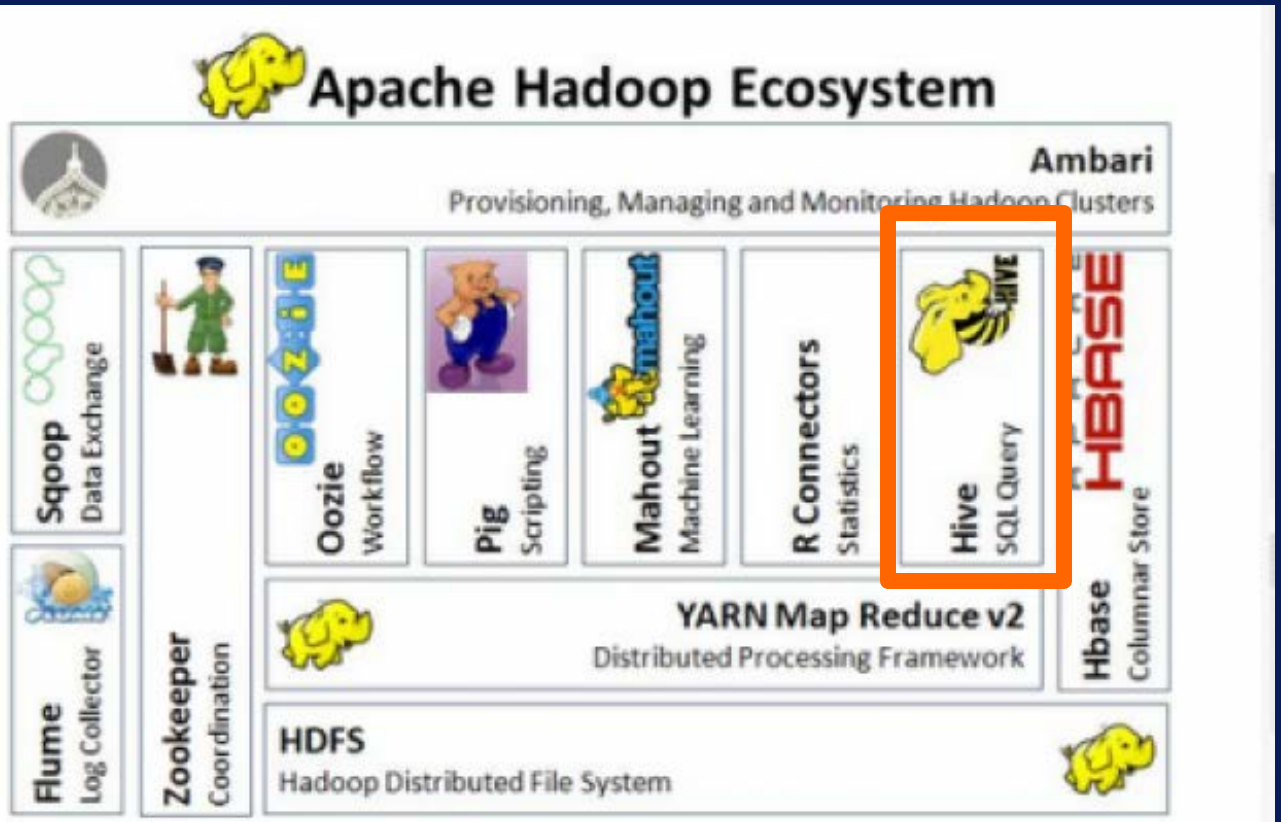

ETL Tool
Extract
Transform
Load
Online
Data Store
Original
Data
Transformed
Data
Hadoop
Data Mart
Query
BI Tools

Pig for ETL


ETL Tool
Extract
Transform
Load
Online
Data Store
riginal
Data
Transformed
Data
Hadoop
Data Mart
Query
BI Tools

Pig for ETL


ETL Tool
Extract
Transform
Load
Online
Data Store
Original
Data
Transformed
Data
Hadoop
Data Mart
Query
BI Tools


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

Apache Hive

- Data warehouse software facilitates querying and managing large datasets residing in distributed storage

Apache Hive

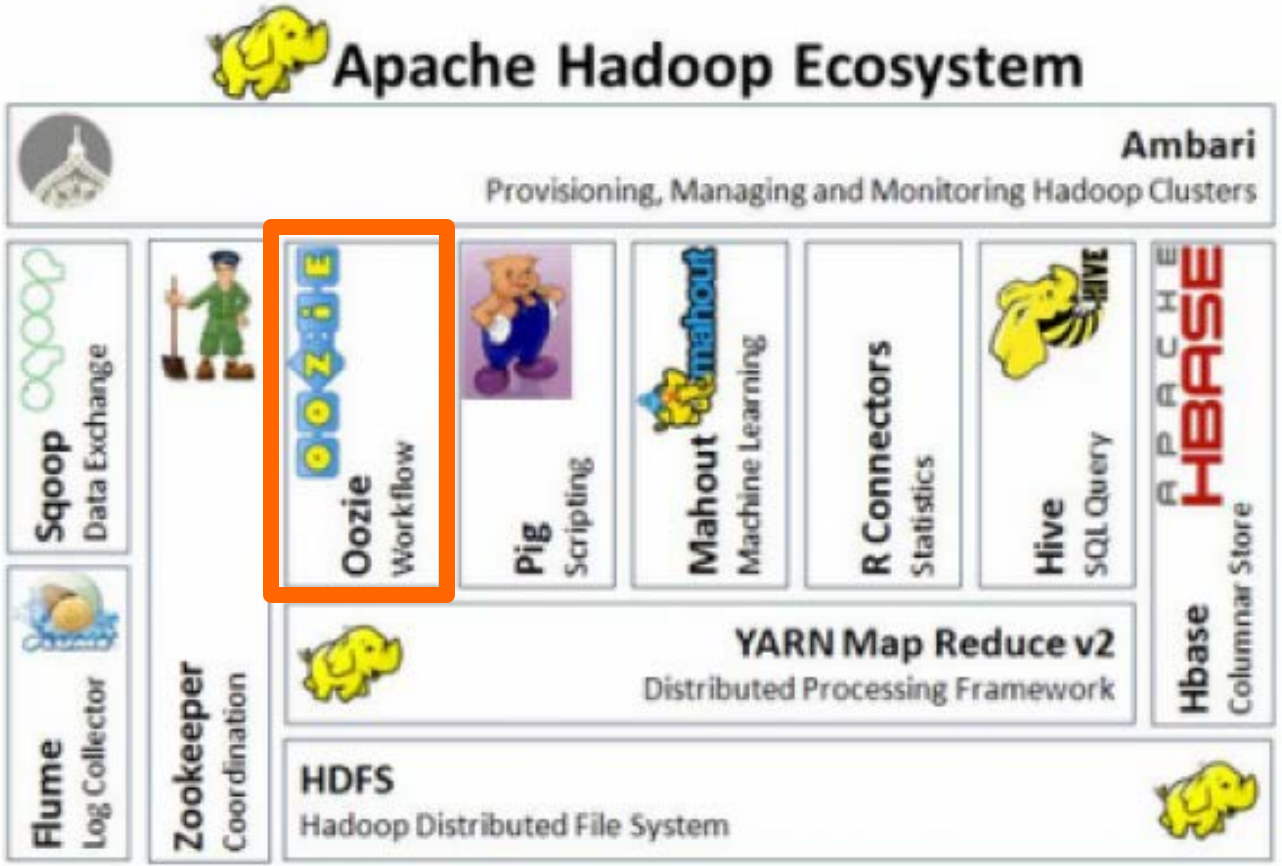
SQL-like language!

Apache Hive

Facilitates querying and managing large datasets in HDFS

Apache Hive

Mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
APACHE HBASE
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

Oozie

Workflow scheduler system to manage Apache Hadoop jobs

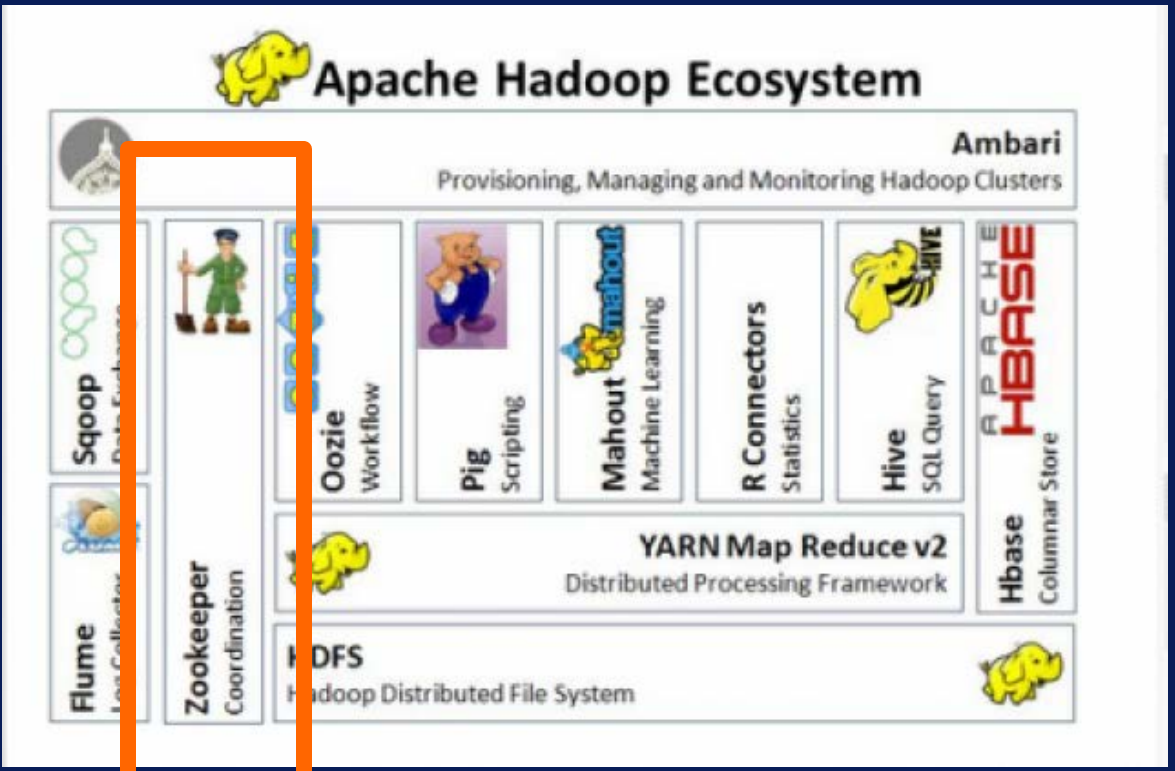
Oozie

Oozie Coordinator jobs!

Oozie

Supports

MapReduce, Pig, Apache Hive, and Sqoop, etc.


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
APACHE
HBASE
Hbase
Columnar Store
Flume
Zookeeper
Coordination
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

Zookeeper

Provides operational services for a Hadoop cluster group services

Zookeeper

Centralized service for:
maintaining configuration information
naming services
providing distributed synchronization
and providing group services

Zookeeper

Centralized service for:

maintaining configuration information

Zookeeper

Centralized service for:

maintaining configuration information

naming services

Zookeeper

Centralized service for:

maintaining configuration information

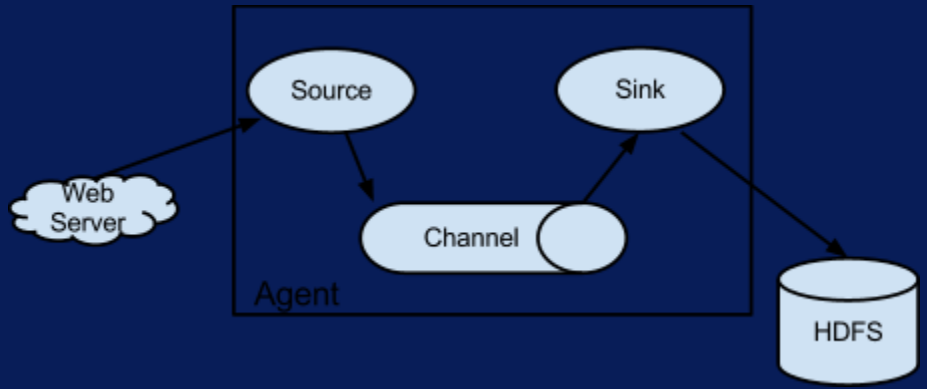
naming services

providing distributed synchronization

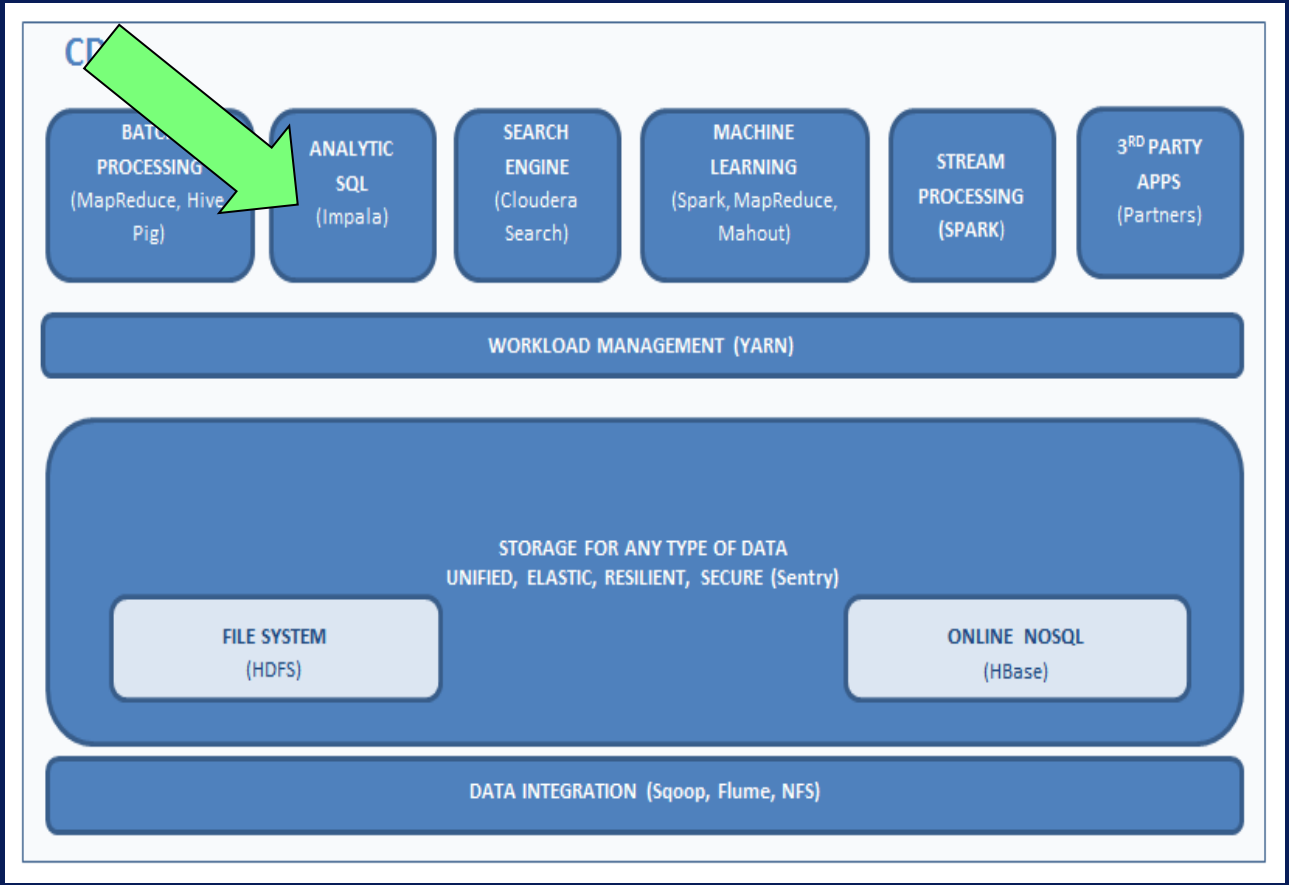
and providing group services

Flume

Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data



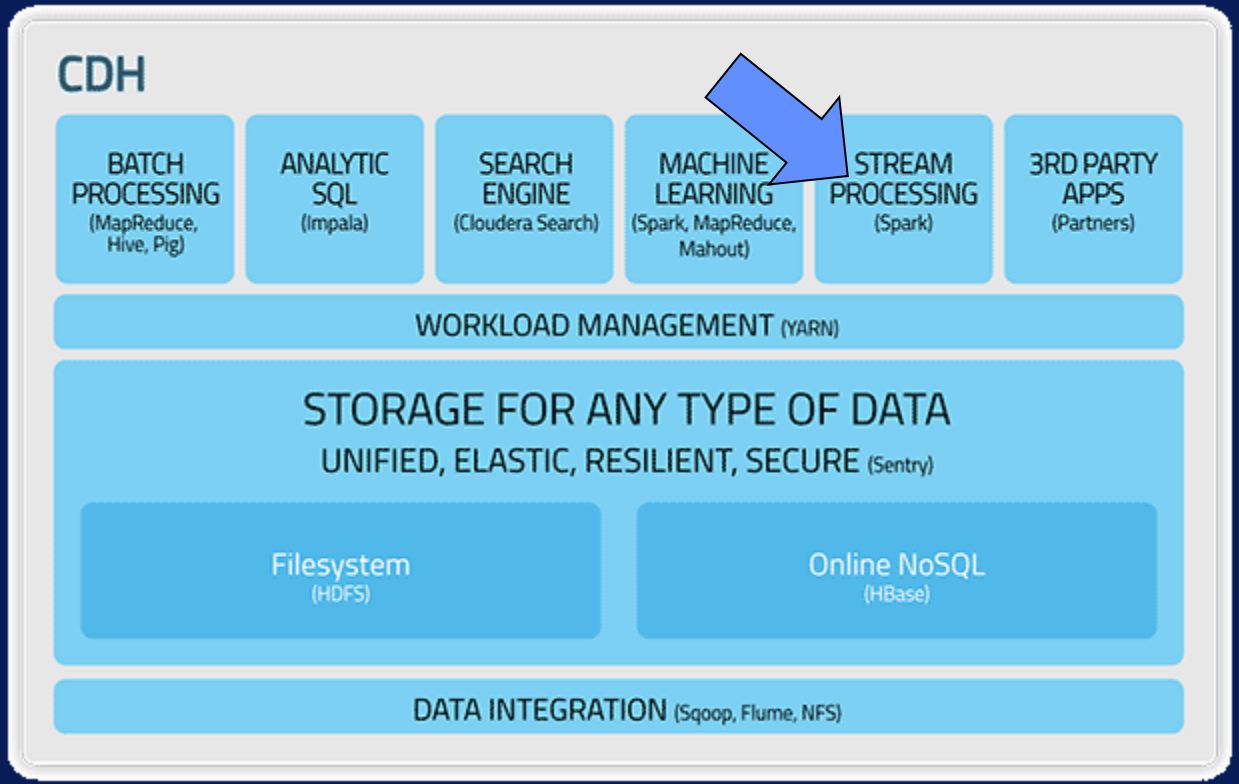
Additional Cloudera Hadoop Components
Impala


CD
BATC
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)
FILE SYSTEM
(HDFS)
ONLINE NOSQL
(HBase)
DATA INTEGRATION (Sqoop, Flume, NFS)

Impala

- Cloudera's open source massively parallel processing (MPP) SQL query engine Apache Hadoop

Additional Cloudera Hadoop Components
Spark
The New Paradigm


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)

Spark

Apache Spark™ is a fast and general engine for large-scale data processing

Spark Benefits

Multi-stage in-memory primitives provides performance up to 100 times faster for certain applications

Spark Benefits

Allows user programs to load data into a cluster's memory and query it repeatedly

- Well-suited to machine learning!!!

Up Next

Tour of the Cloudera’s Quick Start VM