Fundamentals of Big Data Analytics

Lecture 13-14 – Hadoop Ecosystem

Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

Apache Framework Hadoop Modules

Apache Framework Basic Modules

Hadoop Common

Hadoop Distributed File System (HDFS)

Hadoop YARN

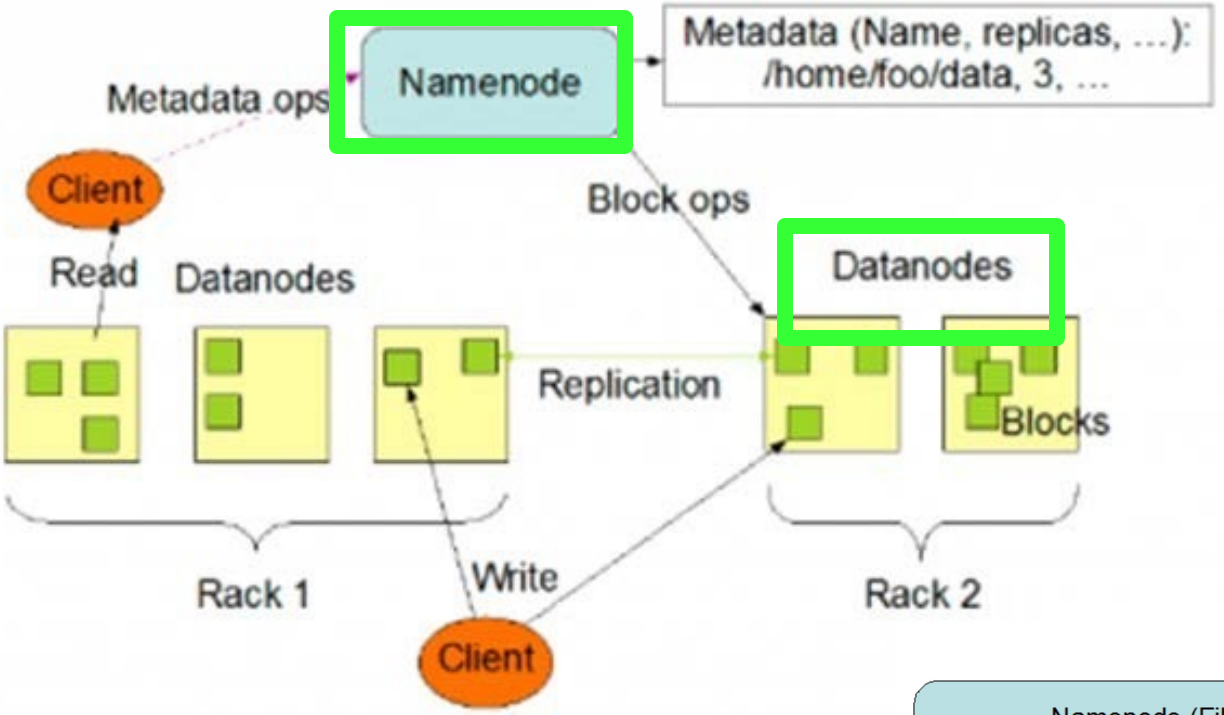
Hadoop MapReduce

HDFS

Hadoop Distributed File System

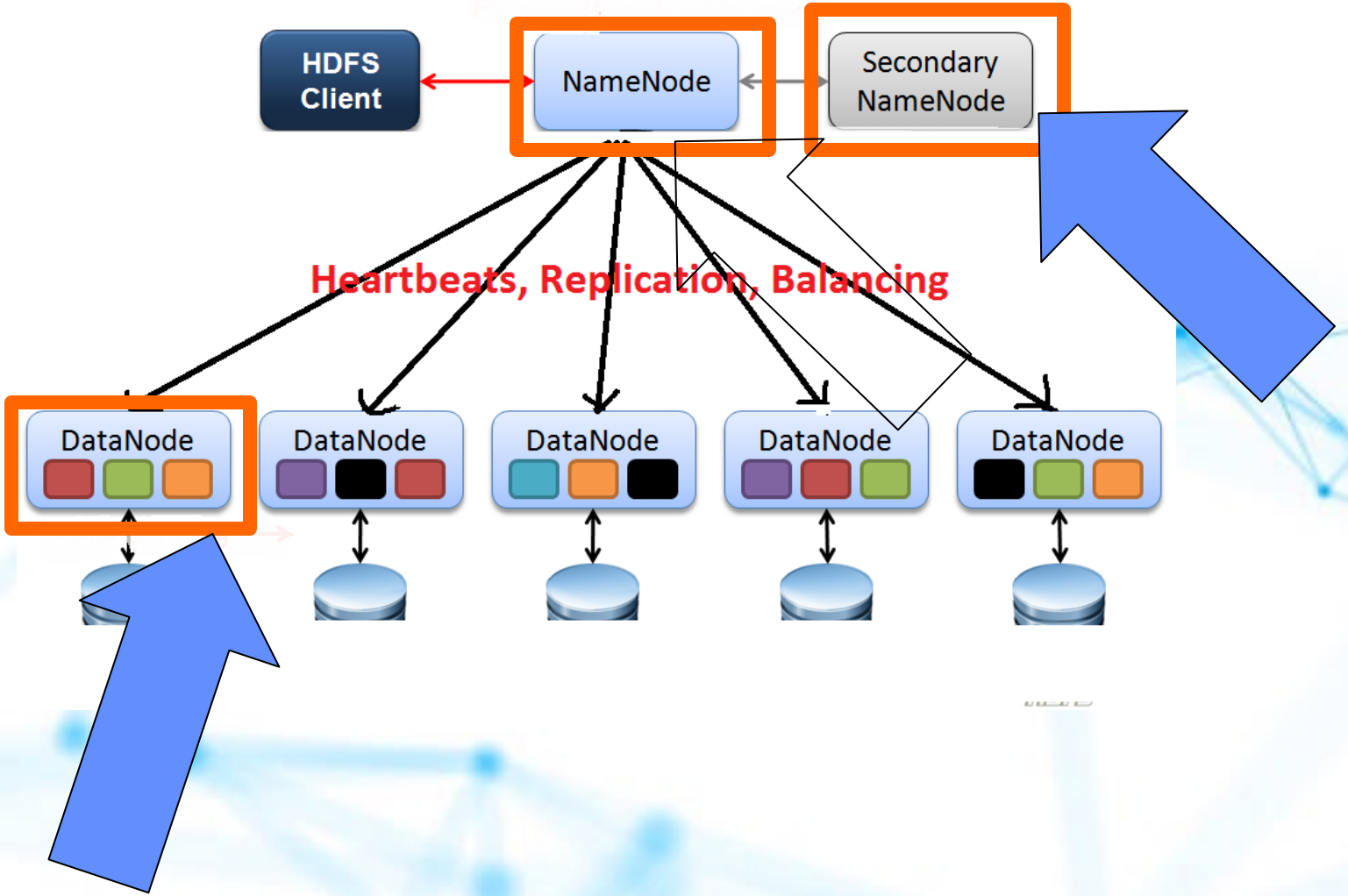
Distributed, scalable, and portable file-system written in Java for the Hadoop framework

HDFS

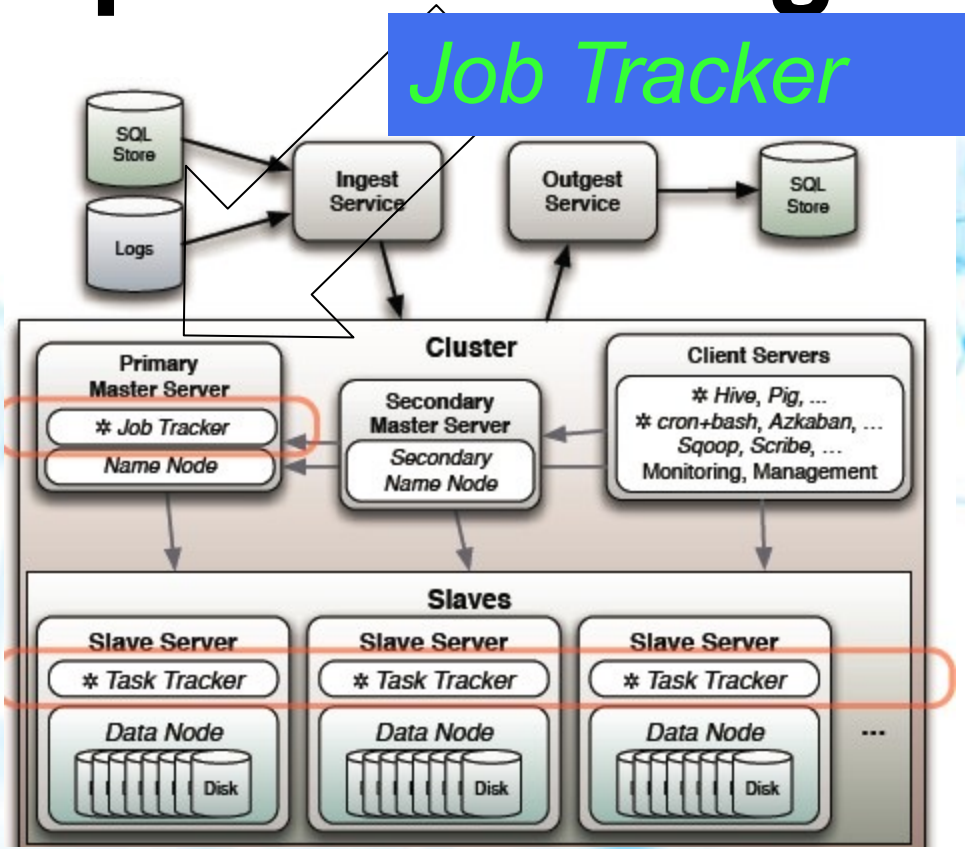


Block Replication

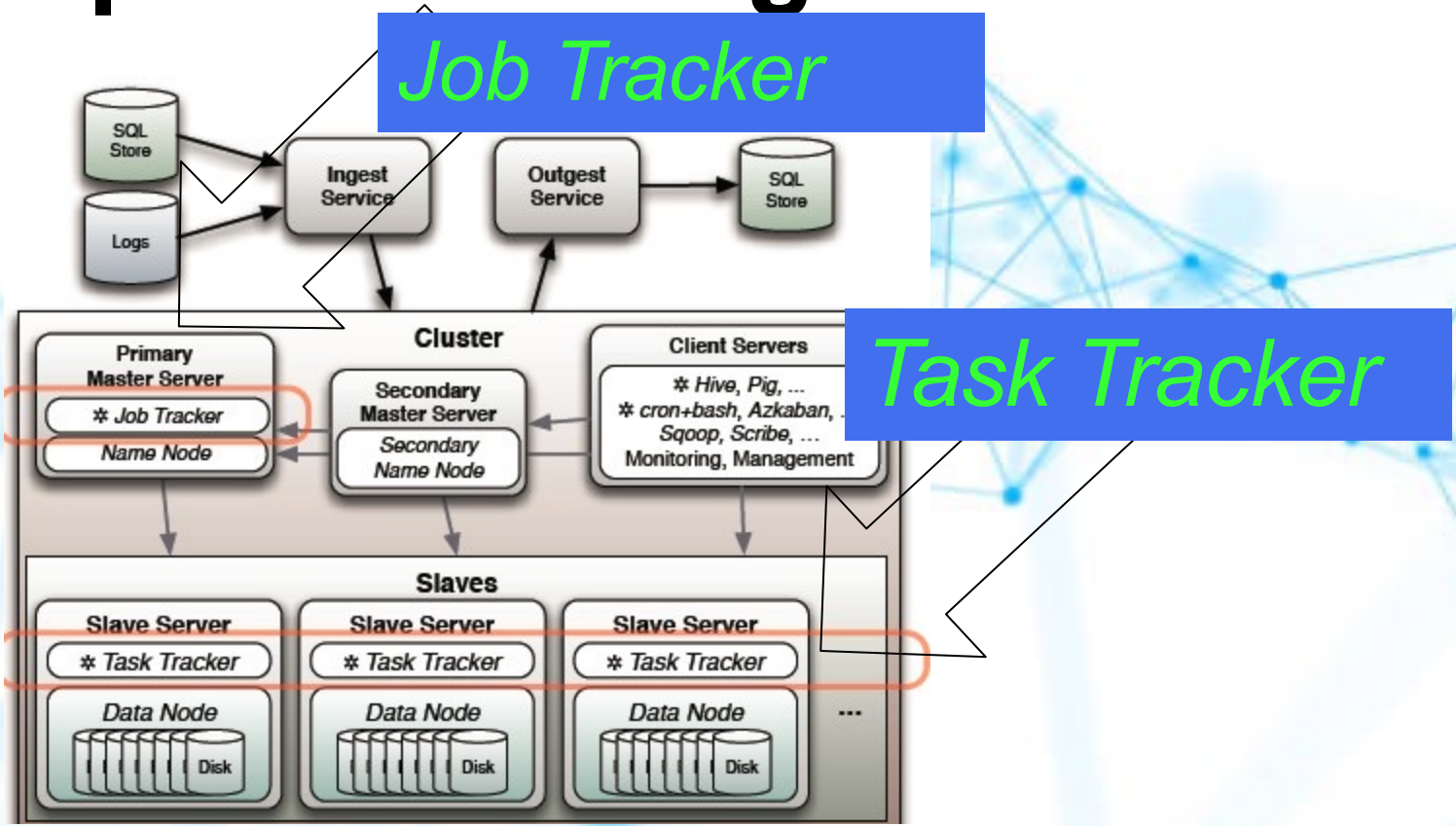
Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, {1,3}, ...
/users/sameerp/data/part-1, r:3, {2,4,5}, ...


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

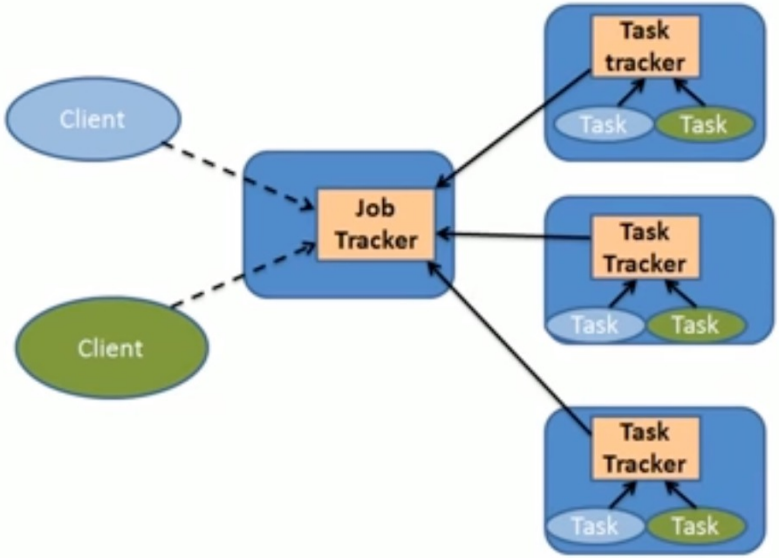
MapReduce Engine



MapReduce Engine

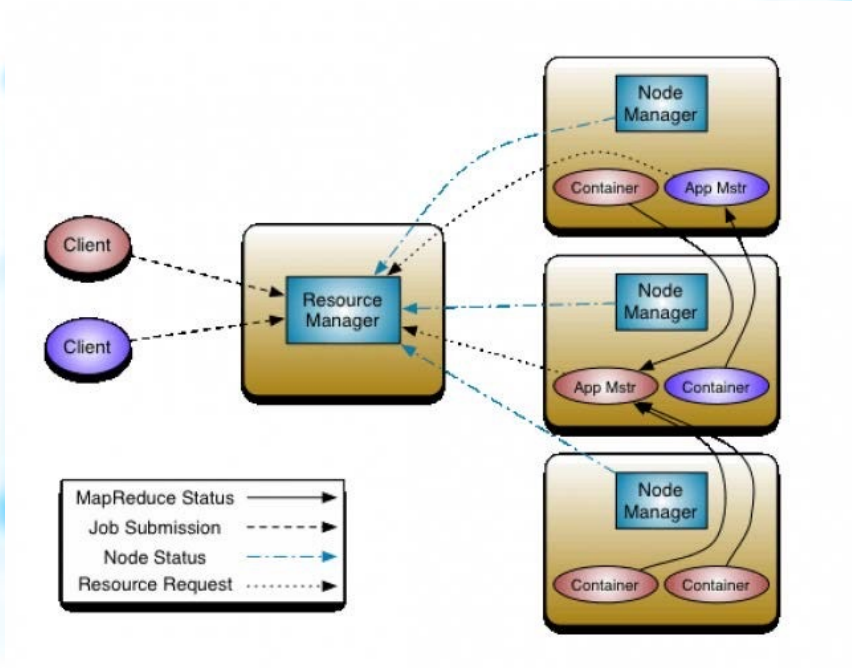


MapReduce Engine

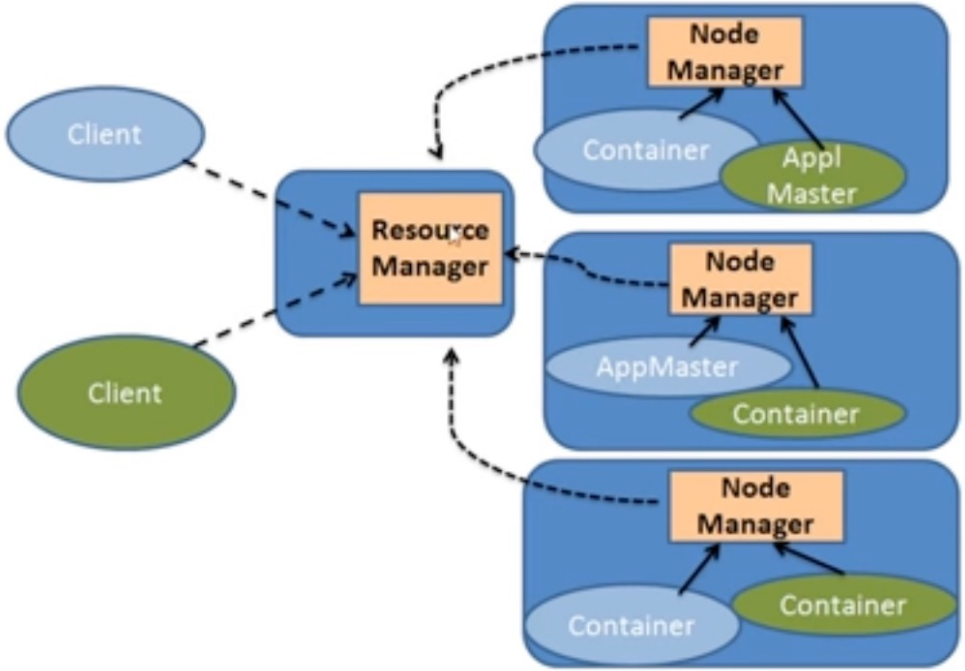


1. JobTracker is a Master daemon
2. Responsible to assign and track task execution progress
3. TaskTrackers are slave daemons
4. They run on systems where data nodes reside
5. Responsible to spawn a child jvm to execute Map, Reduce and intermediate tasks

Apache Hadoop NextGen MapReduce (YARN)



Apache Hadoop NextGen MapReduce (YARN)



✓Job tracker 1.0 responsibility is now split

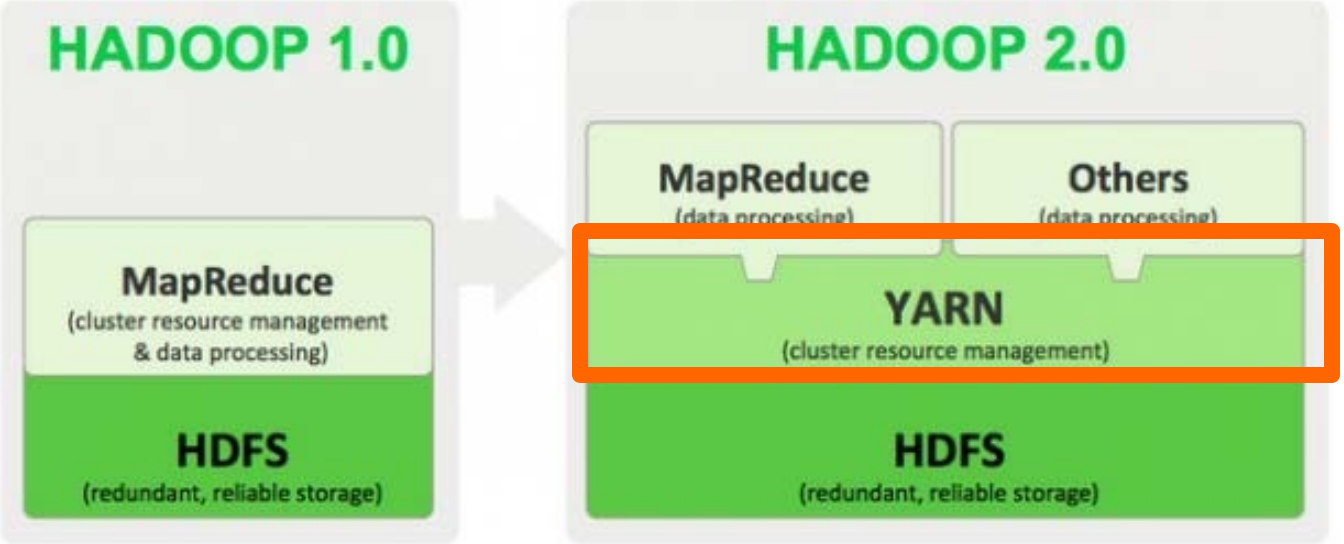
- **Resource Manager** manages the resource allocation in the cluster
- **Application master** manages resource needs of individual applications

✓Node Manager is a generalized task tracker

✓A container executes an application specific process

Apache Hadoop NextGen MapReduce (YARN)

1. **Client**: To submit MapReduce jobs
2. **Resource Manager**: To manage the use of resources across the cluster.
3. **Container** : Name given to a package of resources including RAM, CPU, Network, HDD etc.
4. **Node Manager**: to oversee the containers running on the cluster nodes.
5. **Application Master**: which negotiates with the Resource Manager for resources and runs the application-specific process (Map or Reduce tasks) in those clusters.


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

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

Scalability

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

Scalability

- **Scalability** bottleneck caused by having a single JobTracker. According to Yahoo!, the practical limits of such a design are reached with a cluster of 5,000 nodes and 40,000 tasks running concurrently.
- The computational resources on each slave node are divided by a cluster administrator into a fixed number of map and reduce slots.
- Hadoop was designed to run MapReduce jobs only.

Supports other Workloads

MapReduce Compatibility

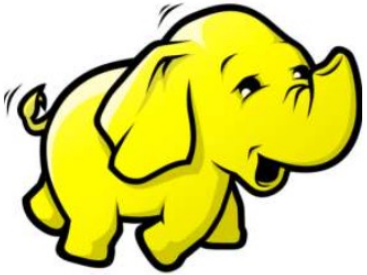
Improved cluster utilization

The Hadoop “Zoo”

Welcome to the Zoo!

Zookeeper

Pig



Hadoop

Jaql

Hama

Shark

Giraph

I am sure you won't find a Shark in any other 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

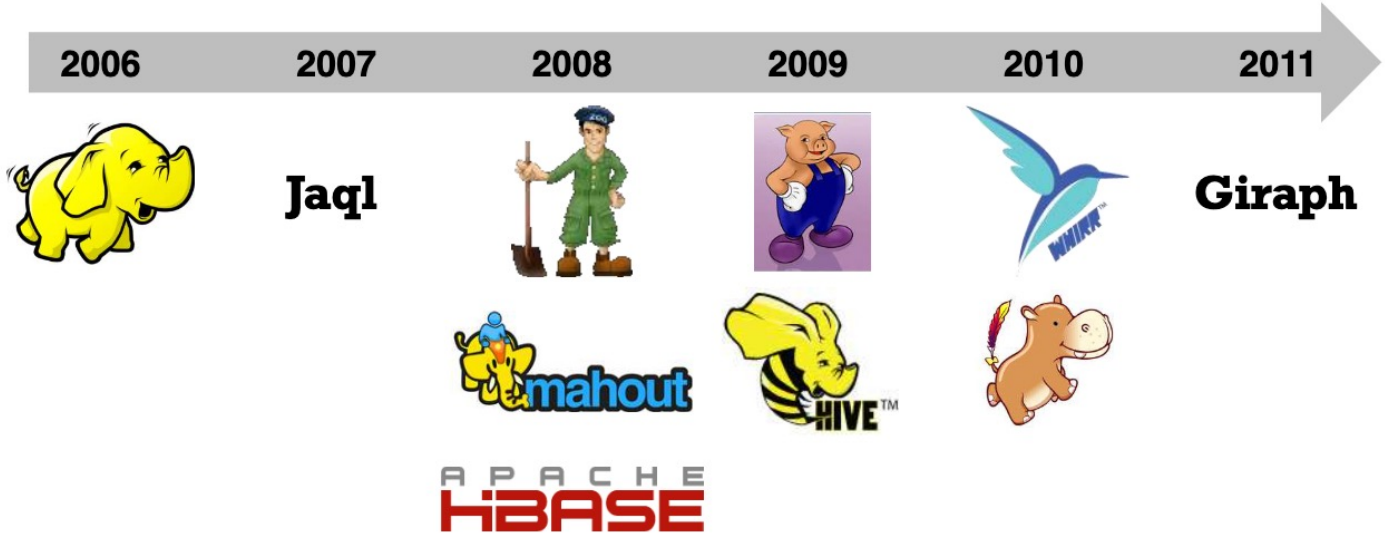
Hbase Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

Evolution Timeline

- **Started by Doug Cutting at Yahoo! in early 2006, and named after his kid's toy elephant**
- **Hadoop committers work at several different organizations**
  - Including Facebook, Yahoo!, LinkedIn, Twitter, Cloudera, Hortonworks



What is Hadoop?

- Hadoop is an open-source project overseen by the Apache Software Foundation
- Hadoop is an ecosystem, not a single product
- Originally based on papers published by Google in 2003 and 2004
- Some of the projects in the ecosystem have been inspired based on whitepapers published by Google

Google calls it:	Hadoop equivalent
GFS	HDFS
MapReduce	Hadoop MapReduce
Sawzall	Hive, Pig
BigTable	HBase
Chubby	ZooKeeper
Pregel	Giraph

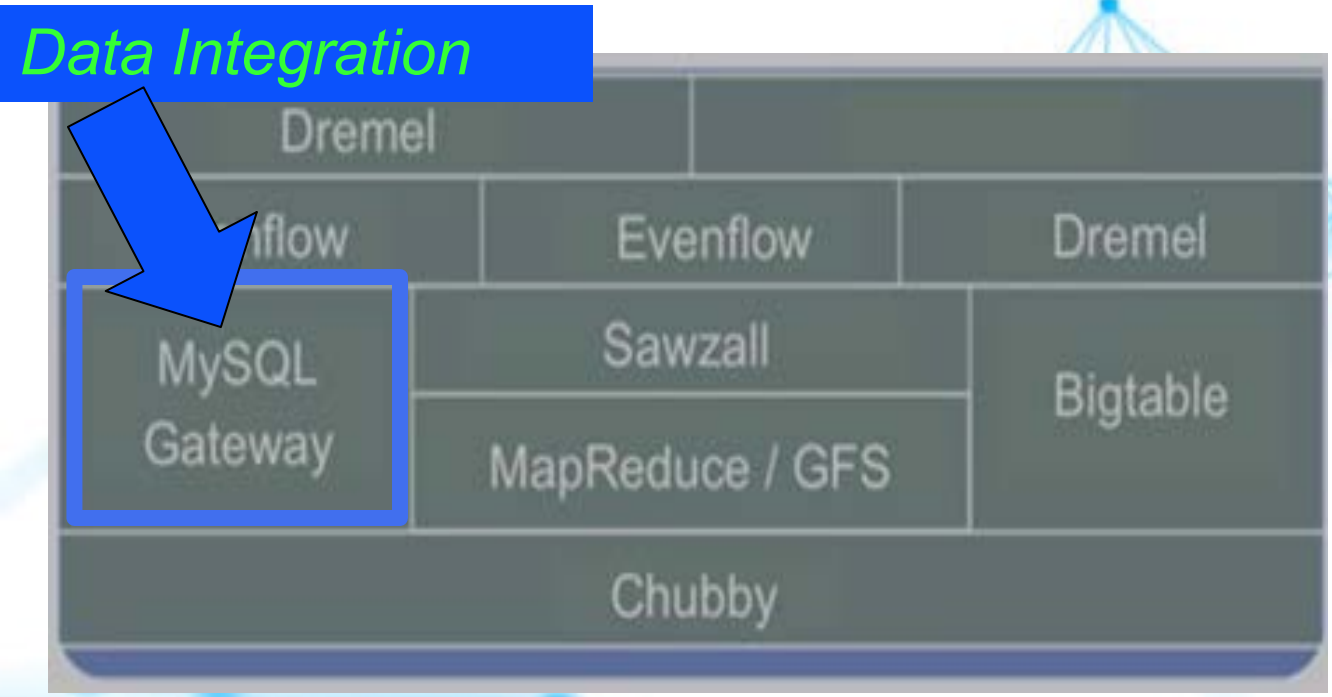
Different Components of Hadoop Ecosystem

HDFS: Hadoop Distributed File System
YARN: Yet Another Resource Negotiator
MapReduce: Programming based Data Processing
Spark: In-Memory data processing
PIG, HIVE: Query based processing of data services
HBase: NoSQL Database
Mahout, Spark MLLib: Machine Learning algorithm libraries
Solar, Lucene: Searching and Indexing
Zookeeper: Managing cluster
Oozie: Job Scheduling

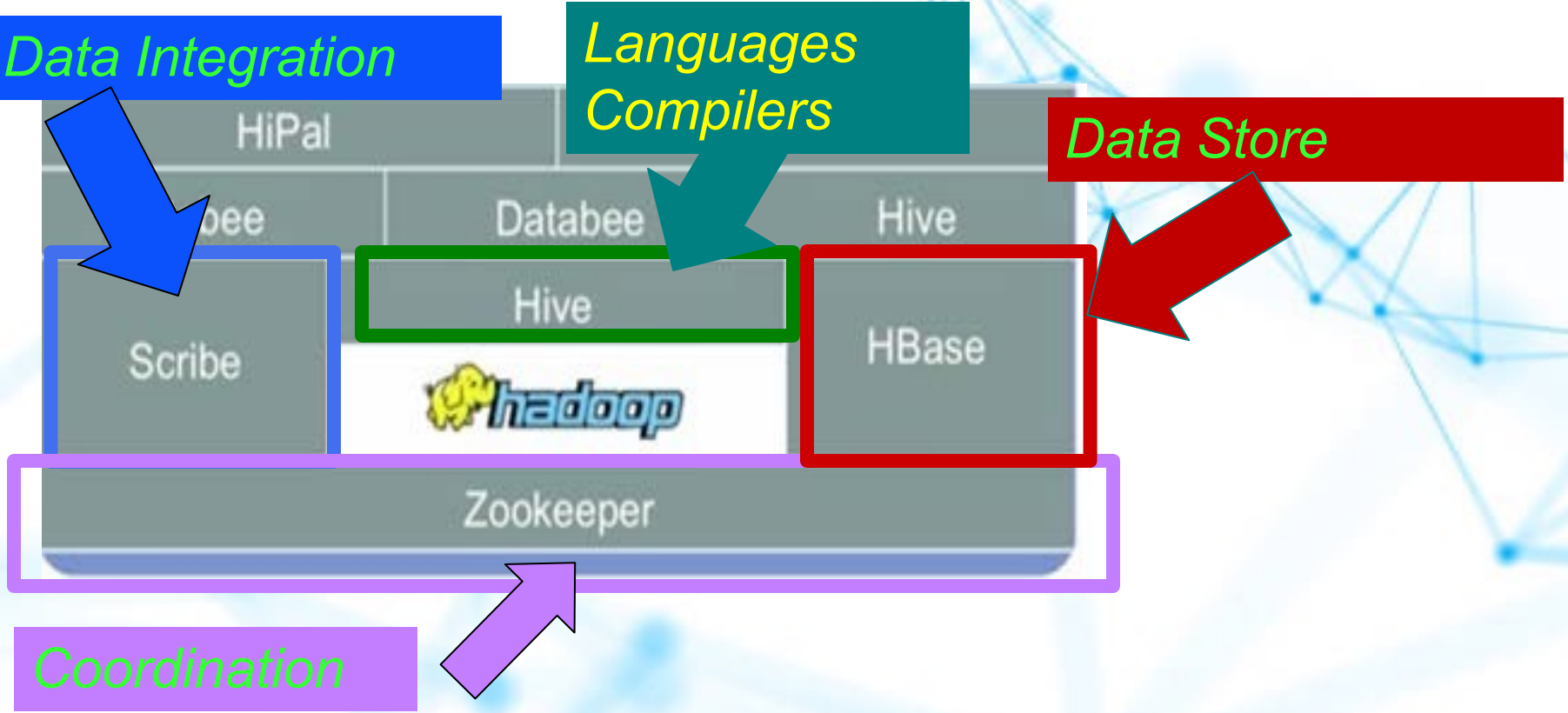
Original Google Stack

Original Google Stack

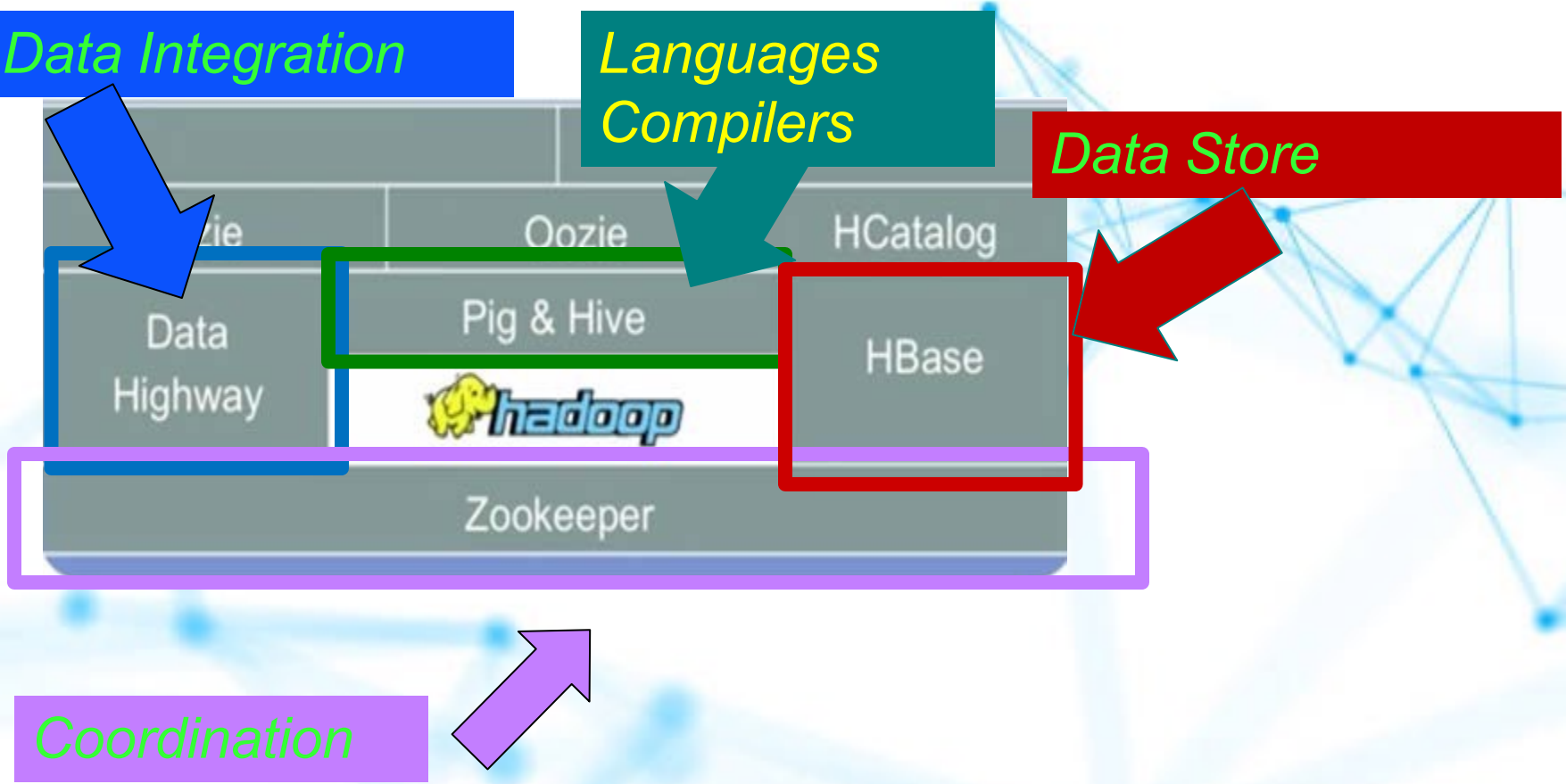
Original Google Stack



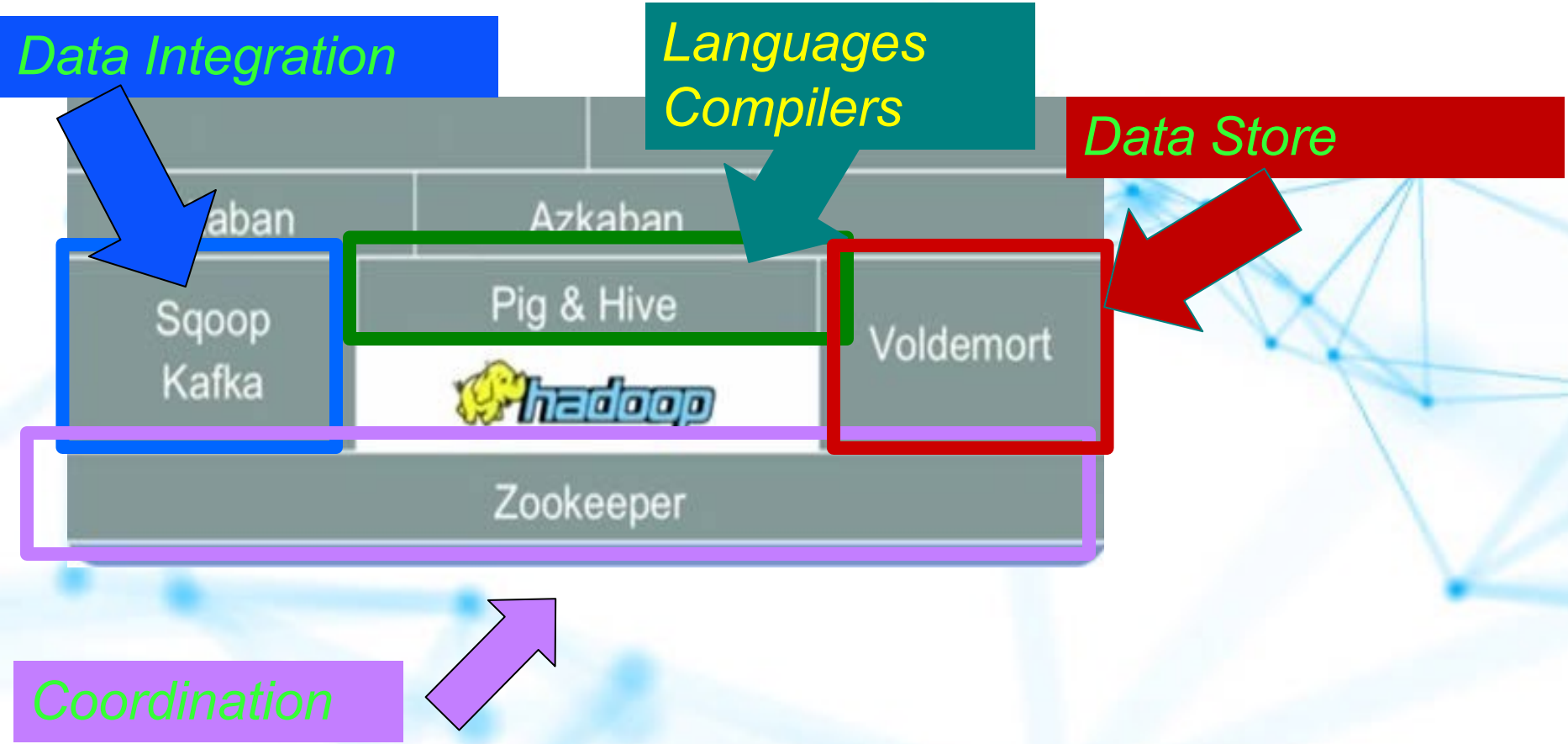
Facebook's Version of the Stack



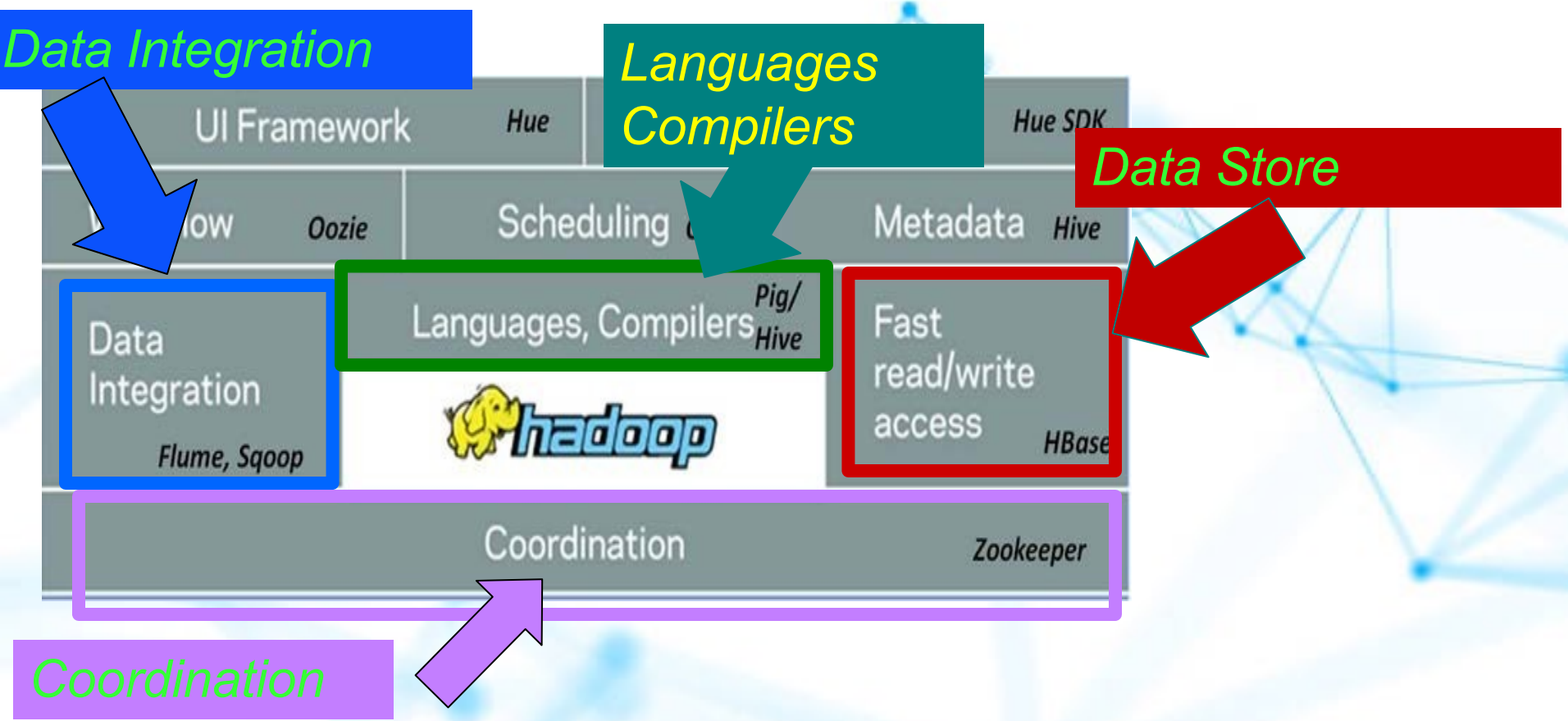
Yahoo's Version of the Stack



LinkedIn's Version of the Stack



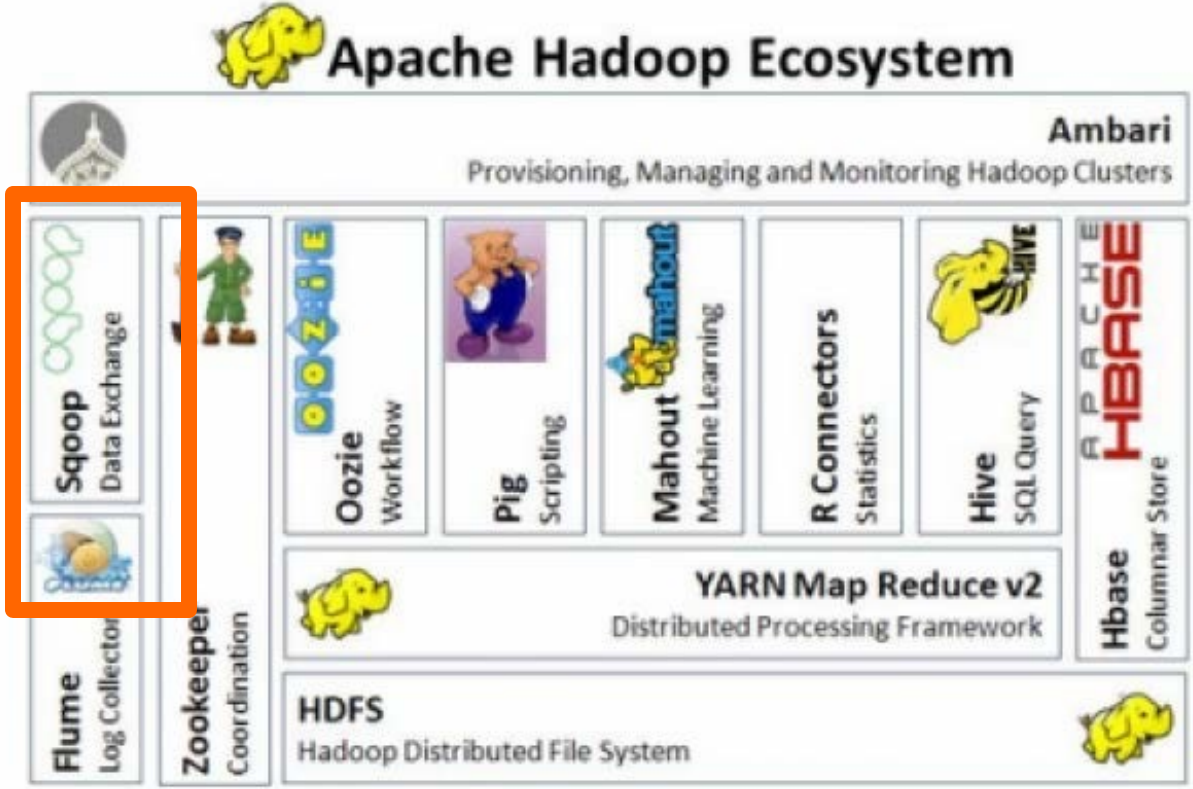
Cloudera's Version of the Stack



Hadoop Ecosystem Major Components

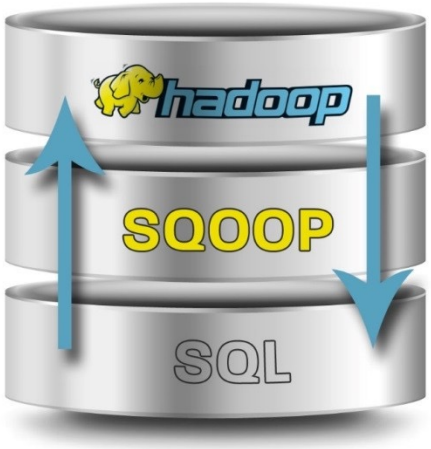
Hadoop Ecosystem Major Components

- *Hadoop Ecosystem* is a platform or a suite which provides various services to solve the big data problems.
- It includes Apache projects and various commercial tools and solutions. There are *four major elements of Hadoop* i.e. **HDFS, MapReduce, YARN, and Hadoop Common**.
- Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

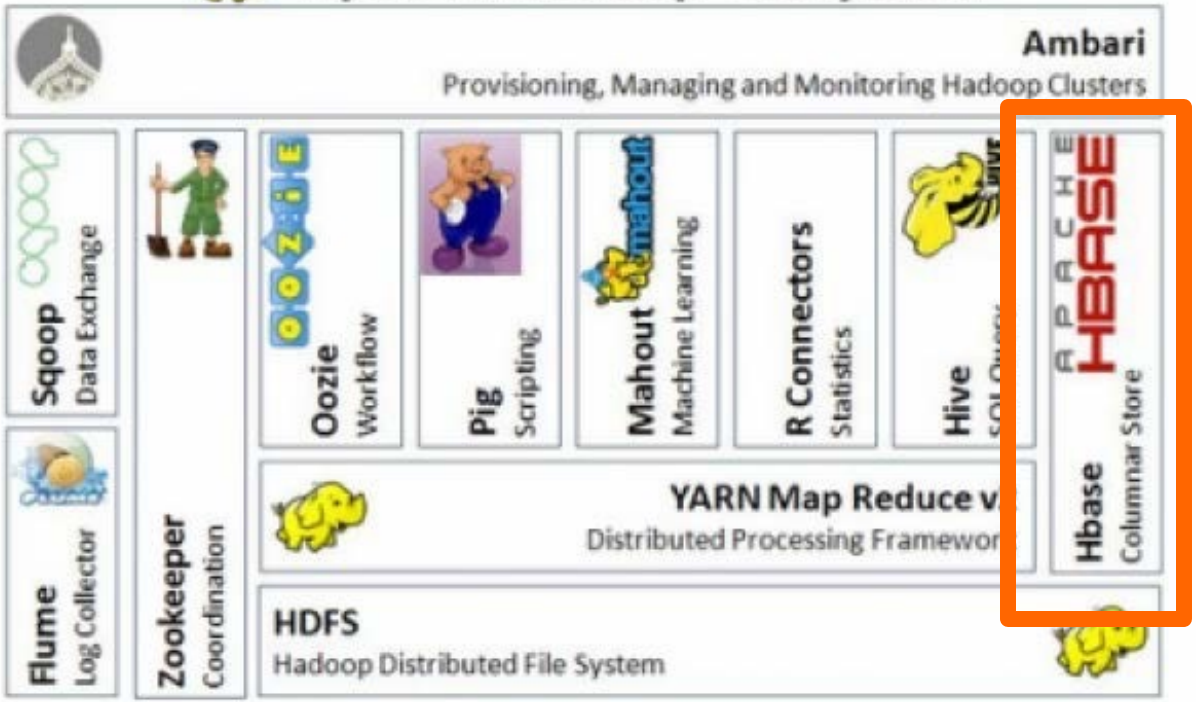

Apache Hadoop Ecosystem
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
YARN Map Reduce v2
Distributed Processing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



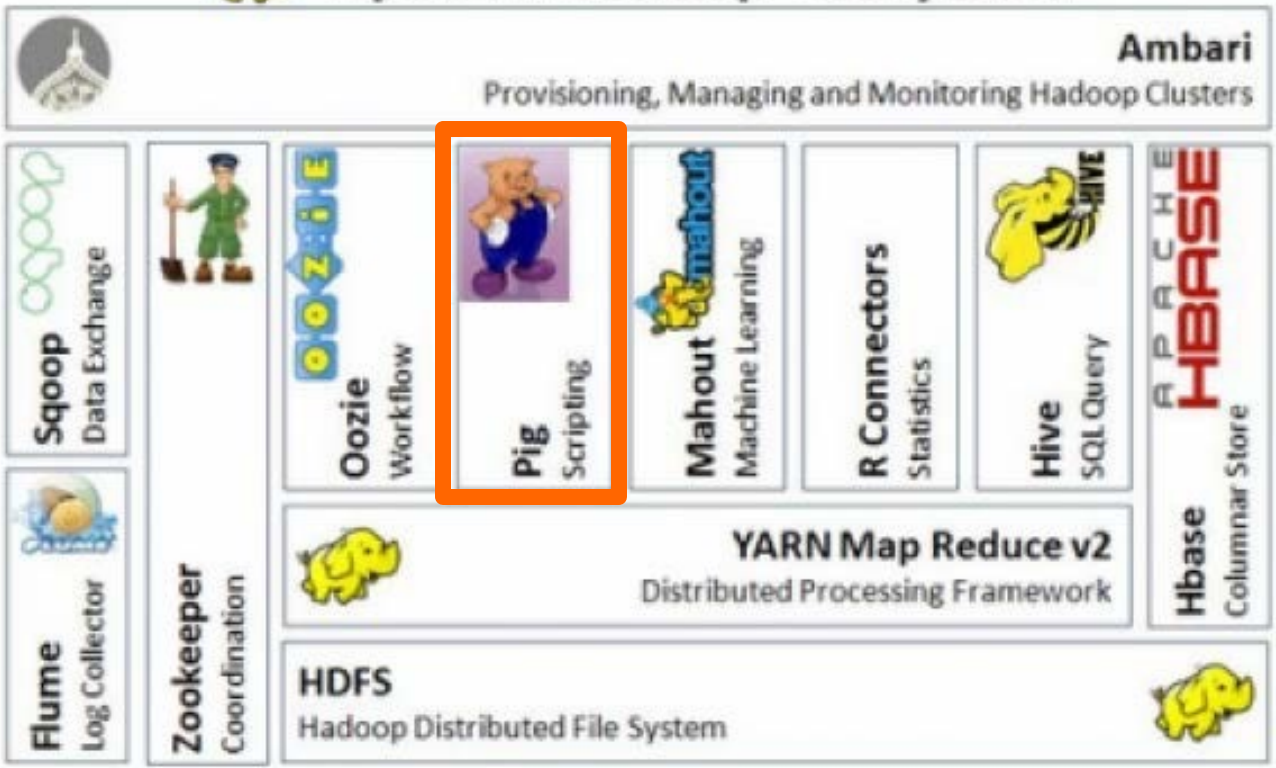
Apache Hadoop Ecosystem



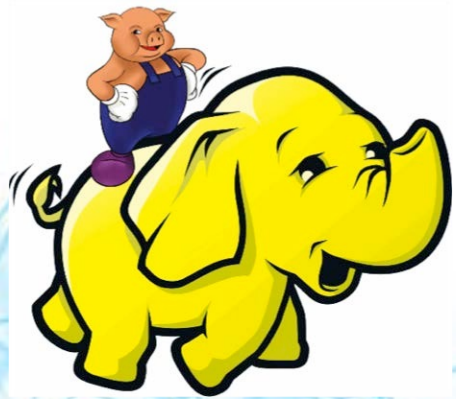
HBASE

- 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

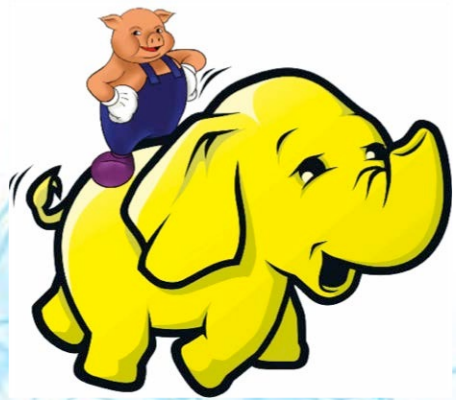


PIG



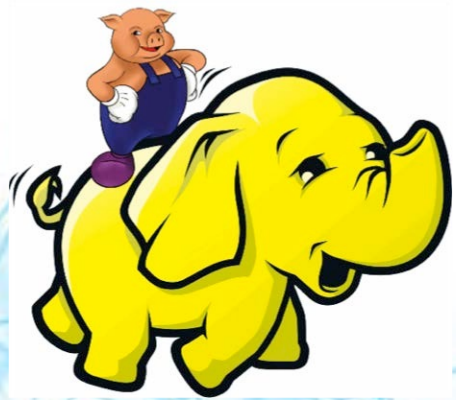
High level programming on top of Hadoop MapReduce

PIG



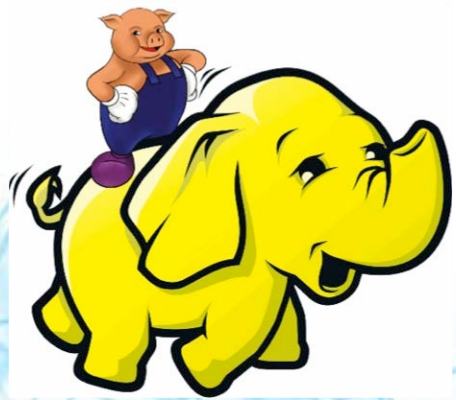
The language: Pig Latin

PIG



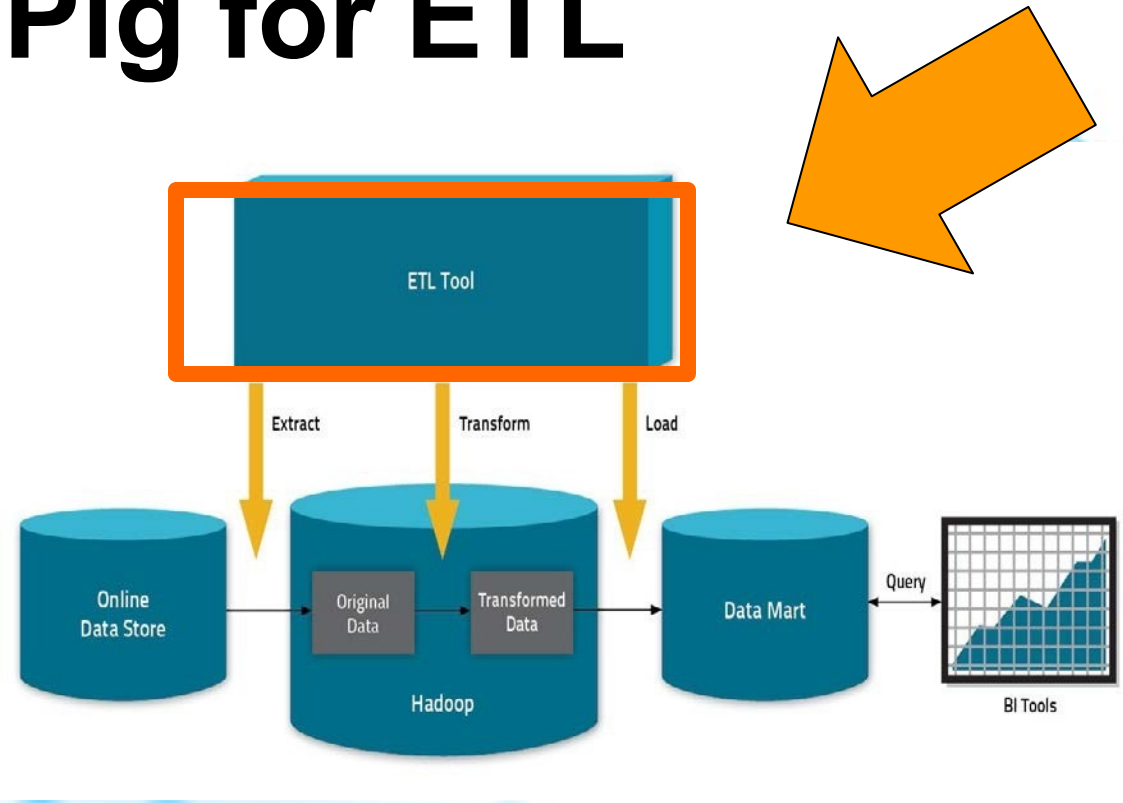
Data analysis problems as data flows

PIG

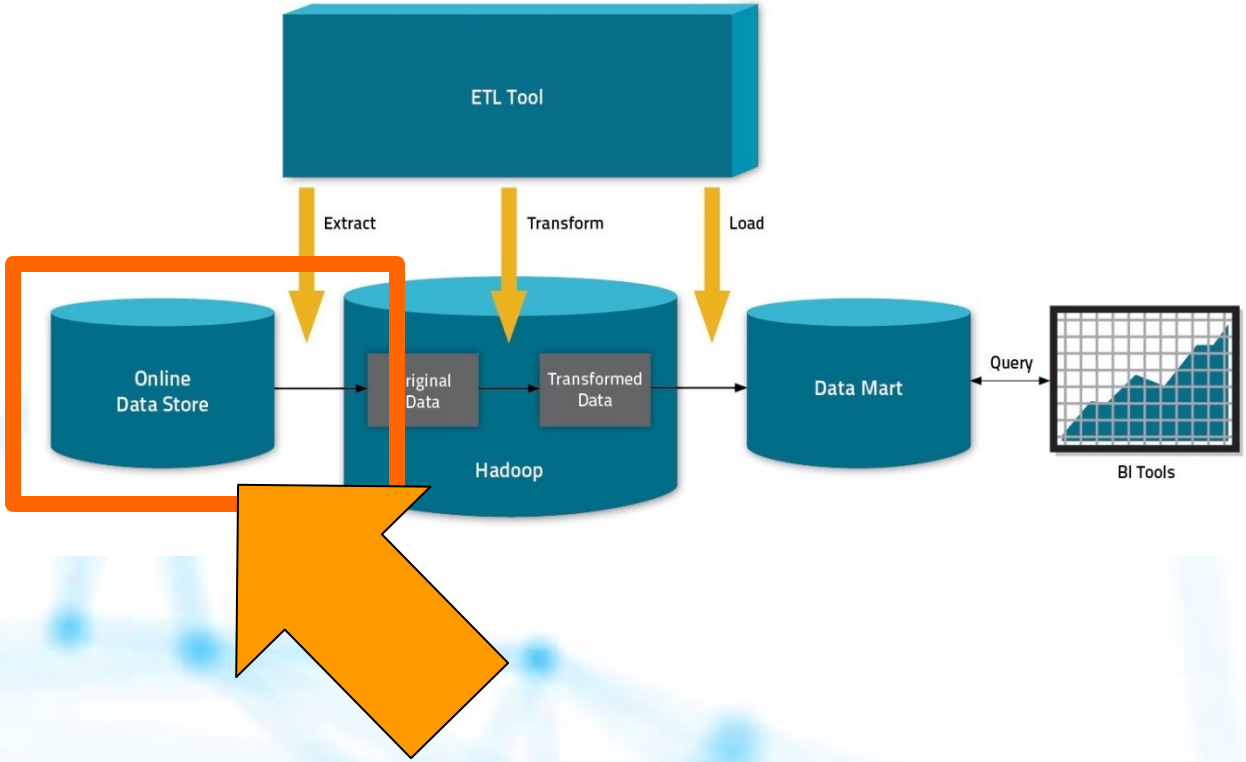


Originally developed at Yahoo 2006

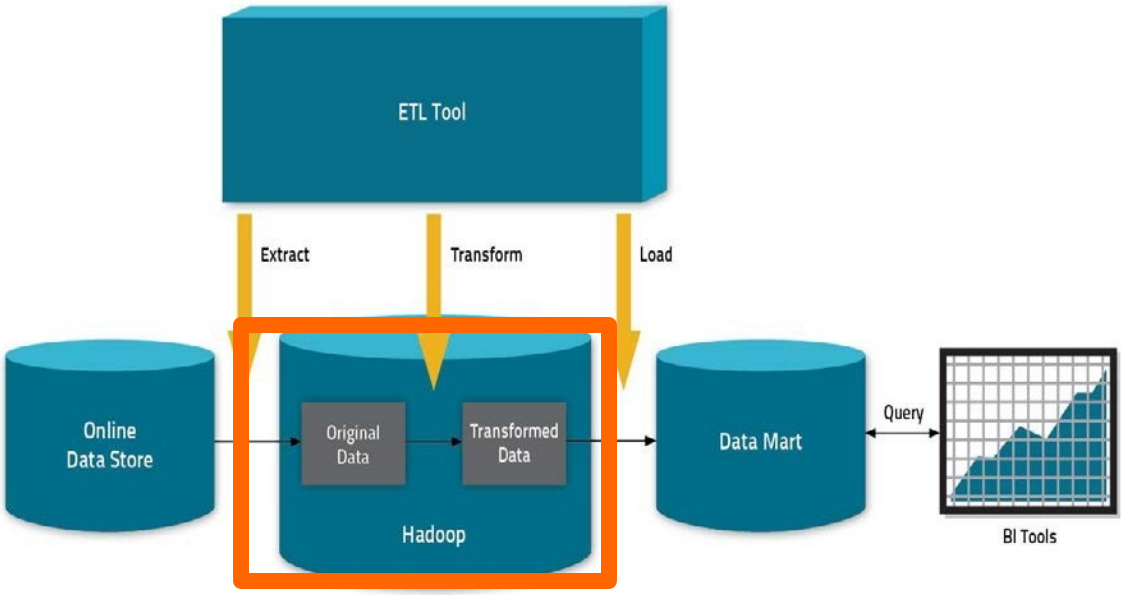
Pig for ETL

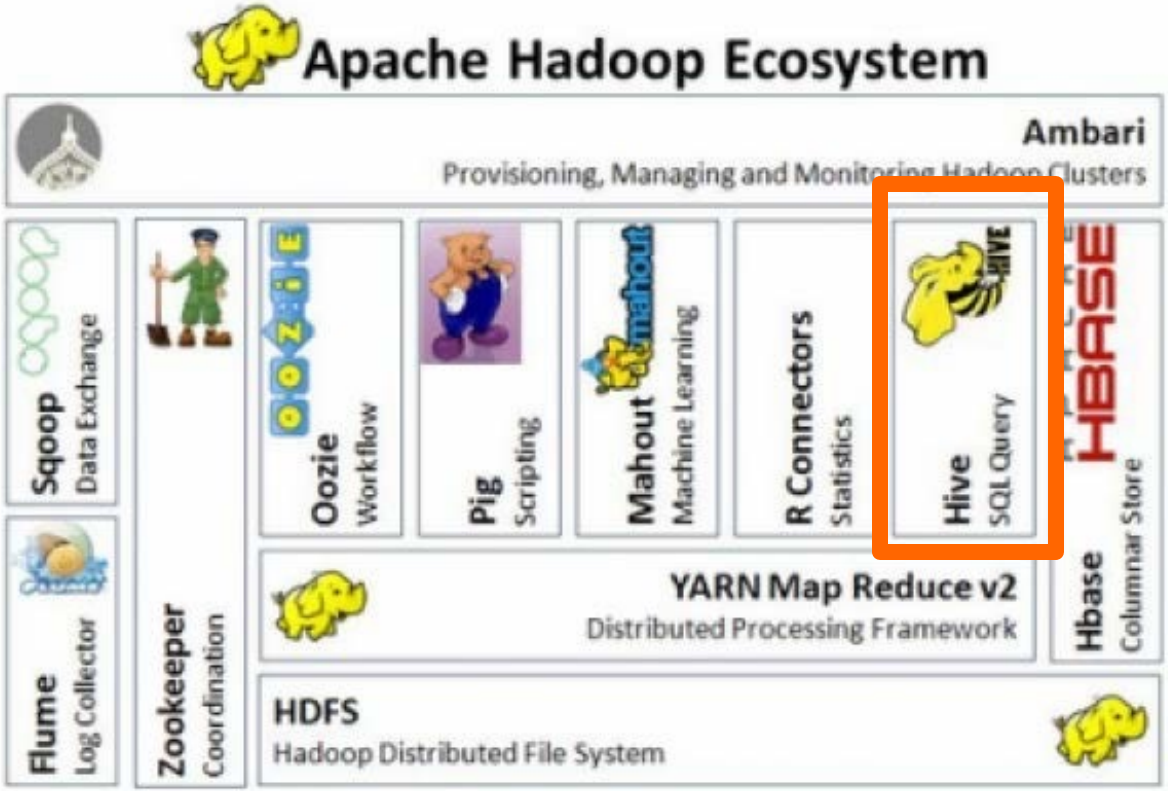

ETL Tool
Extract
Transform
Load
Online
Data Store
Original
Data
Transformed
Data
Hadoop
Data Mart
Query
BI Tools

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
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
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Apache Hive

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

Apache Hive

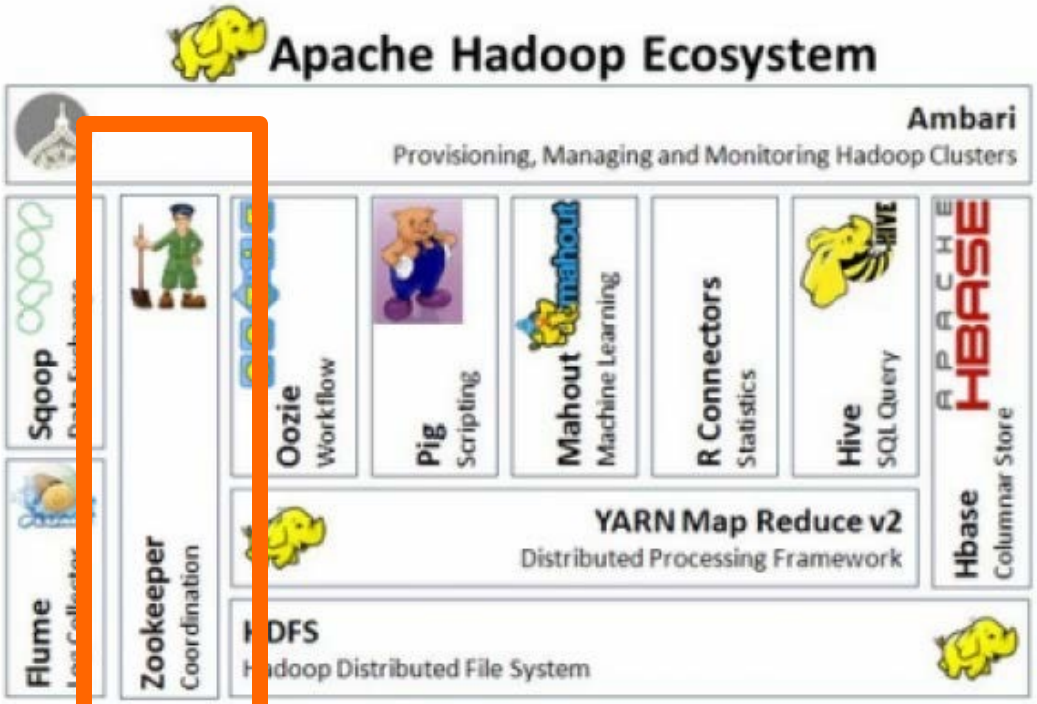
SQL-like language!

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
Sqoop
Flume
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
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

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

Flume

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

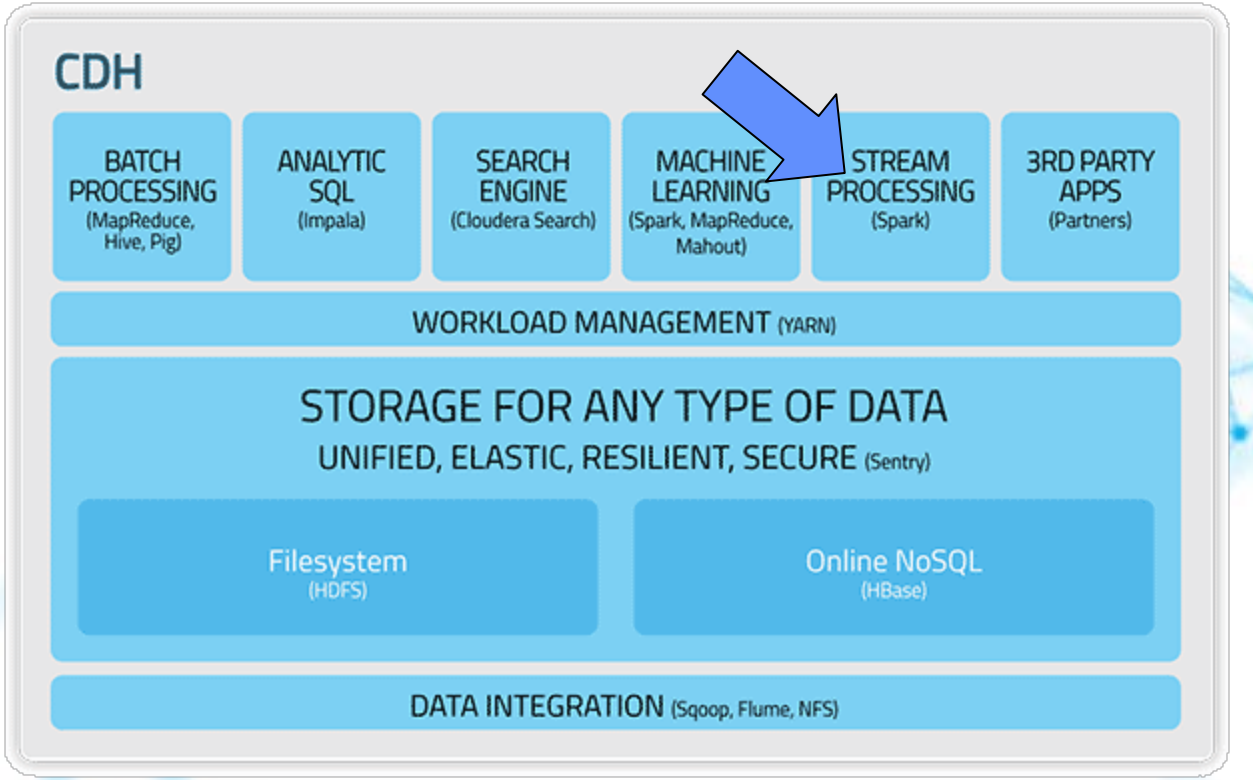
Additional Cloudera Hadoop Components
Impala


CD
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)
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!!!