Hadoop Ecosystem and its components

www.edupristine.com
April 23rd, 2015

Big Data is the buzz word circulating in IT industry from 2008. The amount of data being generated by social networks, manufacturing, retail, stocks, telecom, insurance, banking, and health care industries is way beyond our imaginations.

Before the advent of Hadoop, storage and processing of big data was a big challenge. But now that Hadoop is available, companies have realized the business impact of Big Data and how understanding this data will drive the growth. For example:

• Banking sectors have a better chance to understand loyal customers, loan defaulters and fraud transactions.

• Retail sectors now have enough data to forecast demand.

• Manufacturing sectors need not depend on the costly mechanisms for quality testing. Capturing sensors data and analyzing it would reveal many patterns.

• E-Commerce, social networks can personalize the pages based on customer interests.

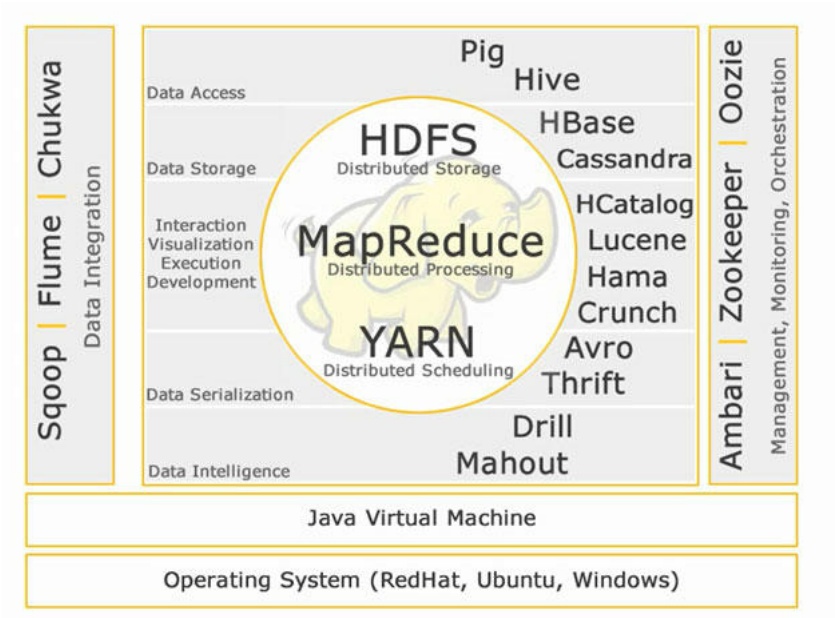
• Stock markets generate humongous amount of data, correlating from time to time will reveal beautiful insights.

Big Data has many useful and insightful applications.

Hadoop is the straight answer for processing Big Data. Hadoop ecosystem is a combination of technologies which have proficient advantage in solving business problems.

Let us understand the components in Hadoop Ecosytem to build right solutions for a given business problem.

Hadoop Ecosystem:



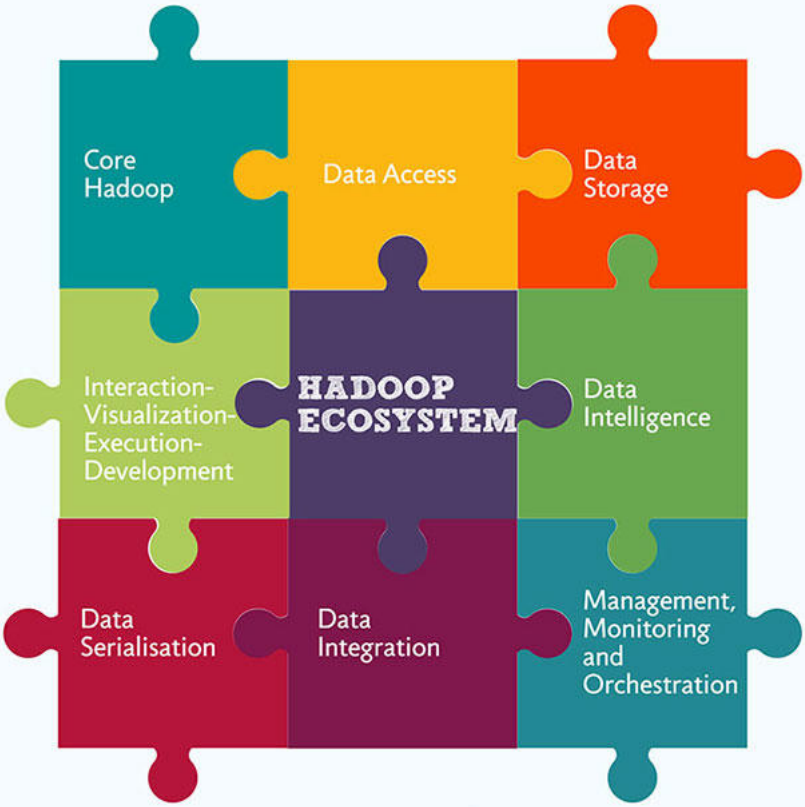
HADOOP ECOSYSTEM

Big Data Handling

Before the advent of Hadoop, storage and processing of big data was a big challenge. But now that Hadoop is available, companies have realized the business impact of Big Data and how understanding this data will drive the growth.

COMPONENTS OF A HADOOP ecosystem

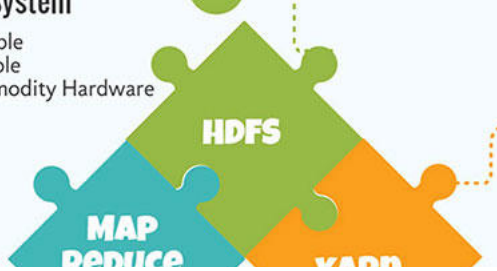
Hadoop is the straight answer for processing Big Data. Hadoop ecosystem has a combination of technologies which have proficient advantage in solving business problems. Let us look at its components



CORE HADOOP

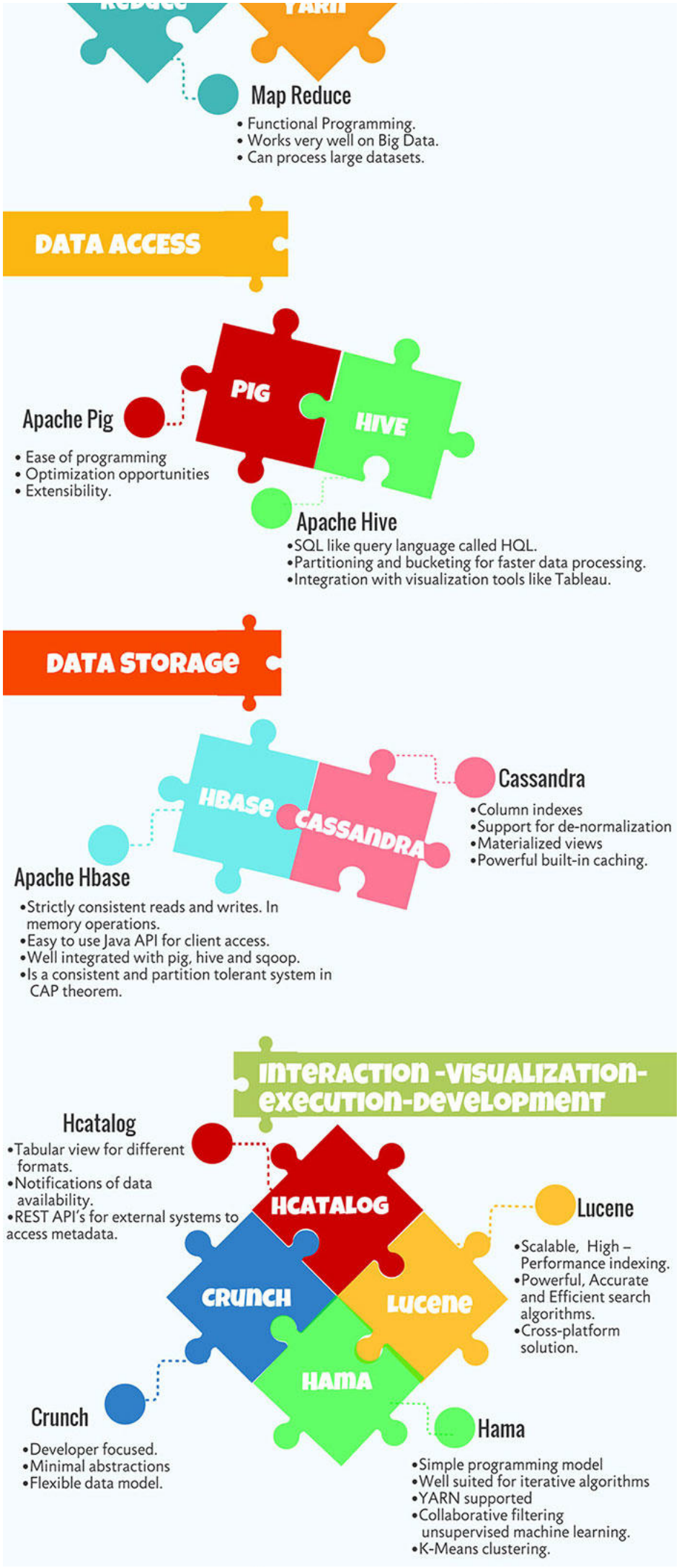
Hadoop Distributed File System

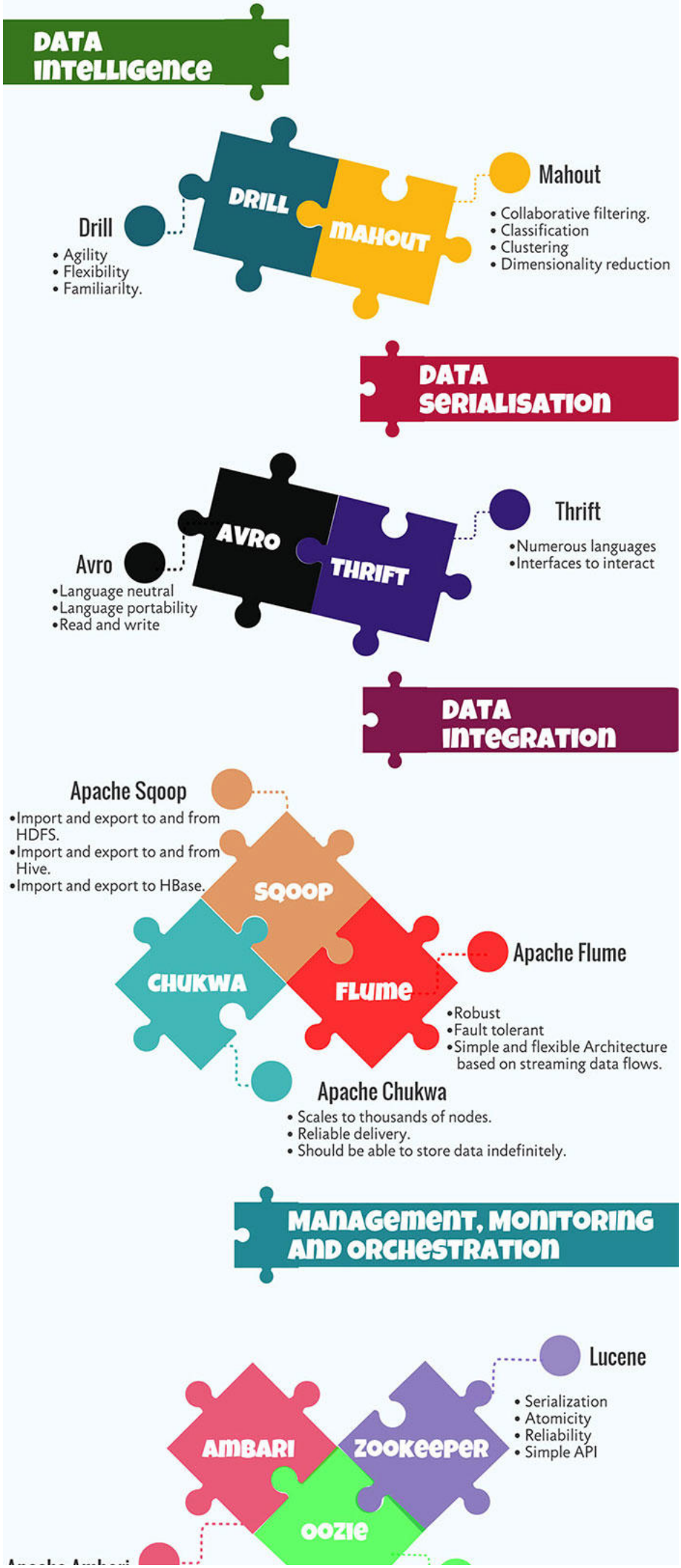
- Scalable
- Reliable
- Commodity Hardware

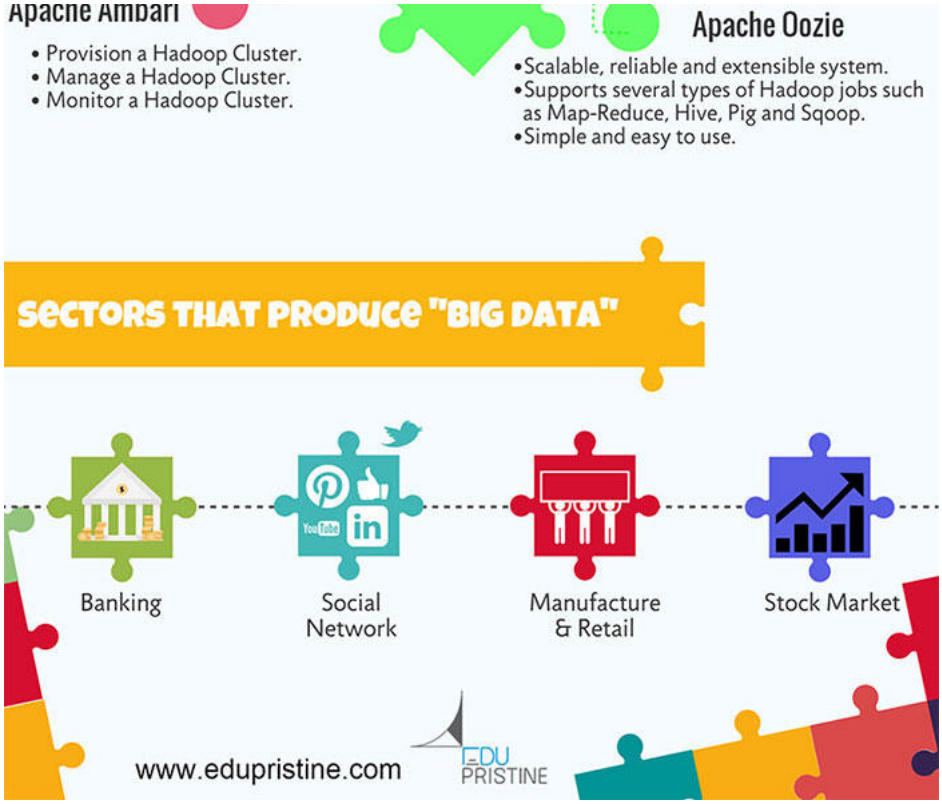


Yet Another Resource Negotiator

- Better resource management.
- Scalability
- Dynamic allocation of cluster resources.


Map Reduce
• Functional Programming.
• Works very well on Big Data.
• Can process large datasets.
DATA ACCESS
PIG
HIVE
Apache Pig
• Ease of programming
• Optimization opportunities
• Extensibility.
Apache Hive
•SQL like query language called HQL.
•Partitioning and bucketing for faster data processing.
•Integration with visualization tools like Tableau.
DATA STORAGE
HBASE
CASSANDRA
Cassandra
•Column indexes
•Support for de-normalization
•Materialized views
•Powerful built-in caching.
Apache Hbase
•Strictly consistent reads and writes. In memory operations.
•Easy to use Java API for client access.
•Well integrated with pig, hive and sqoop.
•Is a consistent and partition tolerant system in CAP theorem.
INTERACTION -VISUALIZATION-EXECUTION-DEVELOPMENT
Hcatalog
•Tabular view for different formats.
•Notifications of data availability.
•REST API's for external systems to access metadata.
HCATALOG
CRUNCH
LUCENE
HAMA
Lucene
•Scalable, High – Performance indexing.
•Powerful, Accurate and Efficient search algorithms.
•Cross-platform solution.
Crunch
•Developer focused.
•Minimal abstractions
•Flexible data model.
Hama
•Simple programming model
•Well suited for iterative algorithms
•YARN supported
•Collaborative filtering unsupervised machine learning.
•K-Means clustering.


DATA INTELLIGENCE
DRILL
MAHOUT
Drill
• Agility
• Flexibility
• Familiarilty.
Mahout
• Collaborative filtering.
• Classification
• Clustering
• Dimensionality reduction
DATA SERIALISATION
AVRO
THRIFT
Avro
•Language neutral
•Language portability
•Read and write
Thrift
•Numerous languages
•Interfaces to interact
DATA INTEGRATION
Apache Sqoop
•Import and export to and from HDFS.
•Import and export to and from Hive.
•Import and export to HBase.
SQOOP
CHUKWA
FLUME
Apache Flume
•Robust
•Fault tolerant
•Simple and flexible Architecture based on streaming data flows.
Apache Chukwa
• Scales to thousands of nodes.
• Reliable delivery.
• Should be able to store data indefinitely.
MANAGEMENT, MONITORING AND ORCHESTRATION
AMBARI
ZOOKEEPER
OOZIE
Lucene
• Serialization
• Atomicity
• Reliability
• Simple API



Core Hadoop:

HDFS:

HDFS stands for Hadoop Distributed File System for managing big data sets with High Volume, Velocity and Variety. HDFS implements master slave architecture. Master is Name node and slave is data node.

Features:

• Scalable

• Reliable

• Commodity Hardware

HDFS is the well known for Big Data storage.

Map Reduce:

Map Reduce is a programming model designed to process high volume distributed data. Platform is built using Java for better exception handling. Map Reduce includes two deamons, Job tracker and Task Tracker.

Features:

• Functional Programming.

• Works very well on Big Data.

• Can process large datasets.

Map Reduce is the main component known for processing big data.

YARN:

YARN stands for Yet Another Resource Negotiator. It is also called as MapReduce 2(MRv2). The two major functionalities of Job Tracker in MRv1, resource management and job scheduling/ monitoring are split into separate daemons which are ResourceManager, NodeManager and ApplicationMaster.

Features:

• Better resource management.

• Scalability

• Dynamic allocation of cluster resources.

Data Access:

Pig:

Apache Pig is a high level language built on top of MapReduce for analyzing large datasets with simple adhoc data analysis programs. Pig is also known as Data Flow language. It is very well integrated with python. It is initially developed by yahoo.

Salient features of pig:

• Ease of programming

• Optimization opportunities

• Extensibility.

Pig scripts internally will be converted to map reduce programs.

Hive:

Apache Hive is another high level query language and data warehouse infrastructure built on top of Hadoop for providing data summarization, query and analysis. It is initially developed by yahoo and made open source.

Salient features of hive:

• SQL like query language called HQL.

• Partitioning and bucketing for faster data processing.

• Integration with visualization tools like Tableau.

Hive queries internally will be converted to map reduce programs.

If you want to become a big data analyst, these two high level languages are a must know!!

Data Storage:

Hbase:

Apache HBase is a NoSQL database built for hosting large tables with billions of rows and millions of columns on top of Hadoop commodity hardware machines. Use Apache Hbase when you need random, realtime read/write access to your Big Data.

Features:
• Strictly consistent reads and writes. In memory operations.

• Easy to use Java API for client access.

• Well integrated with pig, hive and sqoop.

• Is a consistent and partition tolerant system in CAP theorem.

Cassandra:

Cassandra is a NoSQL database designed for linear scalability and high availability. Cassandra is based on key-value model. Developed by Facebook and known for faster response to queries.

Features:

• Column indexes

• Support for de-normalization

• Materialized views

• Powerful built-in caching.

Interaction -Visualization- execution-development:

Hcatalog:

HCatalog is a table management layer which provides integration of hive metadata for other Hadoop applications. It enables users with different data processing tools like Apache pig, Apache MapReduce and Apache Hive to more easily read and write data.

Features:

• Tabular view for different formats.

• Notifications of data availability.

• REST API's for external systems to access metadata.

Lucene:

Apache LuceneTM is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Features:

• Scalable, High – Performance indexing.

• Powerful, Accurate and Efficient search algorithms.

• Cross-platform solution.

Hama:

Apache Hama is a distributed framework based on Bulk Synchronous Parallel(BSP) computing. Capable and well known for massive scientific computations like matrix, graph and network algorithms.

Features:

• Simple programming model

• Well suited for iterative algorithms

• YARN supported

• Collaborative filtering unsupervised machine learning.

• K-Means clustering.

Crunch:

Apache crunch is built for pipelining MapReduce programs which are simple and efficient. This framework is used for writing, testing and running MapReduce pipelines.

Features:

• Developer focused.

• Minimal abstractions

• Flexible data model.

Data Serialization:

Avro:

Apache Avro is a data serialization framework which is language neutral. Designed for language portability, allowing data to potentially outlive the language to read and write it.

Thrift:

Thrift is a language developed to build interfaces to interact with technologies built on Hadoop. It is used to define and create services for numerous languages.

Data Intelligence:

Drill:

Apache Drill is a low latency SQL query engine for Hadoop and NoSQL.

Features:

• Agility

• Flexibility

• Familiarilty.

Mahout:

Apache Mahout is a scalable machine learning library designed for building predictive analytics on Big Data. Mahout now has implementations apache spark for faster in memory computing.

Features:

• Collaborative filtering.

• Classification

• Clustering

• Dimensionality reduction

Data Integration:

Apache Sqoop:

Apache Sqoop is a tool designed for bulk data transfers between relational databases and Hadoop.

Features:

• Import and export to and from HDFS.

• Import and export to and from Hive.

• Import and export to HBase.

Apache Flume:

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

Features:

• Robust

• Fault tolerant

• Simple and flexible Architecture based on streaming data flows.

Apache Chukwa:

Scalable log collector used for monitoring large distributed files systems.

Features:

• Scales to thousands of nodes.

• Reliable delivery.

• Should be able to store data indefinitely.

Management, Monitoring and Orchestration:

Apache Ambari:

Ambari is designed to make hadoop management simpler by providing an interface for provisioning, managing and monitoring Apache Hadoop Clusters.

Features:

• Provision a Hadoop Cluster.

• Manage a Hadoop Cluster.

• Monitor a Hadoop Cluster.

Zookeeper is a centralized service designed for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

Features:

• Serialization

• Atomicity

• Reliability

• Simple API

Apache Zookeeper:

Apache Oozie:

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Features:

• Scalable, reliable and extensible system.

• Supports several types of Hadoop jobs such as Map-Reduce, Hive, Pig and Sqoop.

• Simple and easy to use.

We will be discussing about the components in detail in the upcoming articles. Stay tuned.