HADOOP AND THEIR ECOSYSTEM

CONTENTS

- History of Hadoop
- What Is Hadoop
- Hadoop Architecture
- Hadoop Services
- Hadoop Ecosystem

  Hdfs, Hive,Hbase,Mapreduce,Pig,Sqoop,Flume, Zookeeper,
- Advantage of Hadoop
- Disadvantage of Hadoop
- Use of Hadoop
- References
- Conclusion

History of hadoop

- Hadoop was created by Doug Cutting who had created the Apache Lucene (Text Search),which is origin in Apache Nutch (Open source search Engine).Hadoop is a part of Apache Lucene Project.Actually Apache Nutch was started in 2002 for working crawler and search
- In January 2008, Hadoop was made its own top-level project at Apache for, confirming success ,By this time, Hadoop was being used by many other companies such as Yahoo!, Facebook, etc.
- In April 2008, Hadoop broke a world record to become the fastest system to sort a terabyte of data.
- Yahoo take test in which To process 1TB of data (1024 columns)

  oracle – 3 ½ day

  teradata – 4 ½ day

  netezza – 2 hour 50 min

  hadoop - 3.4 min

WHAT IS HADOOP

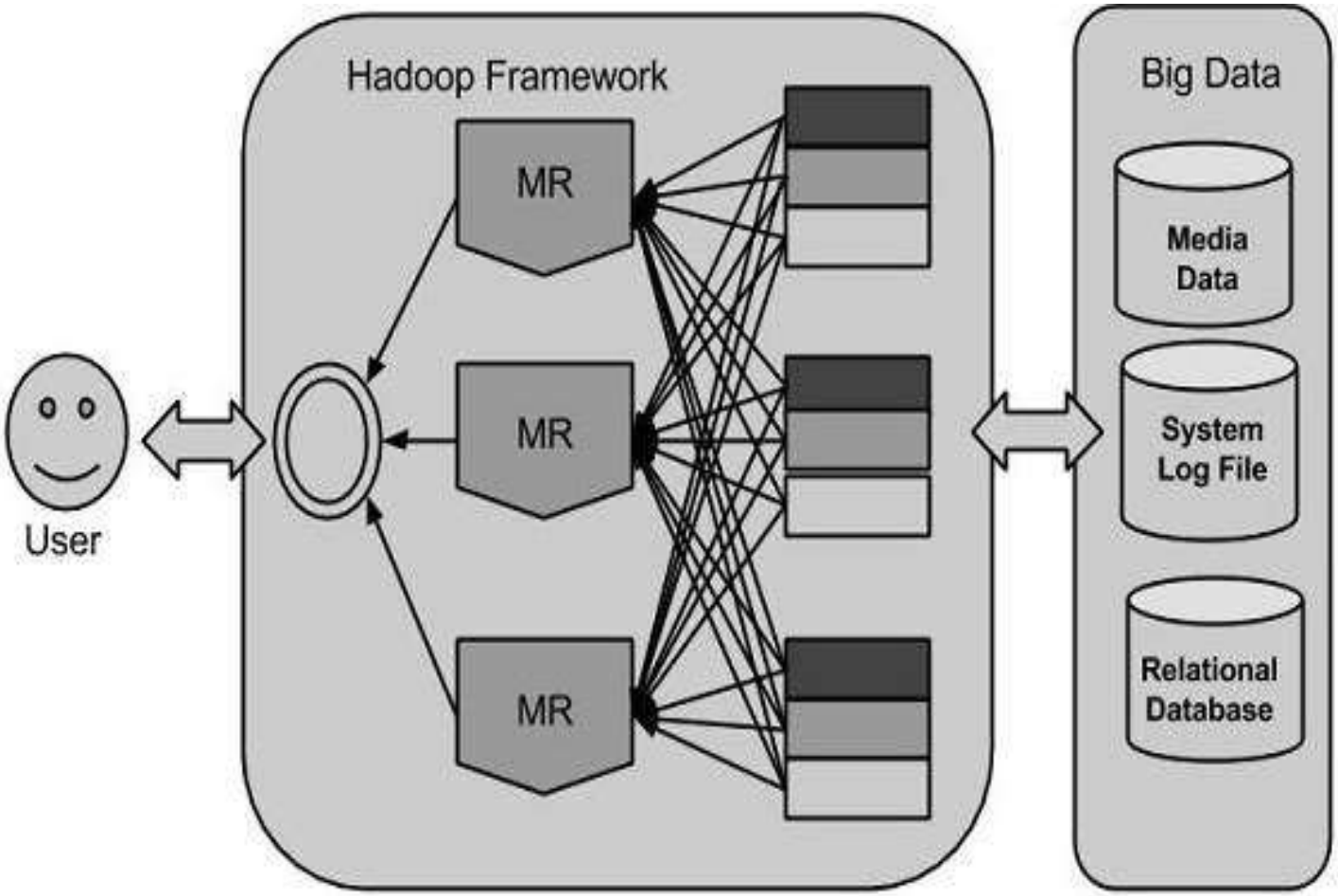
- Hadoop is the product of Apach ,it is the type of distributed system, it is framework for big data
- Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.
- Some of the characteristics:
  - Open source
  - Distributed processing
  - Distributed storage
  - Reliable
  - Economical
  - Flexible

Hadoop Framework Modules

The base Apache Hadoop framework is composed of the following modules:

- **Hadoop Common :**– contains libraries and utilities needed by other Hadoop modules
- **Hadoop Distributed File System (HDFS) :**– a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster
- **Hadoop YARN:**– a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications
- **Hadoop MapReduce:**– an implementation of the MapReduce programming model for large scale data processing.

Framework Architecture



Hadoop Services

- Storage
  1. HDFS (Hadoop distributed file System)

     a)Horizontally Unlimited Scalability

     (No Limit For Max no.of Slaves)

     b)Block Size=64MB(old Version)

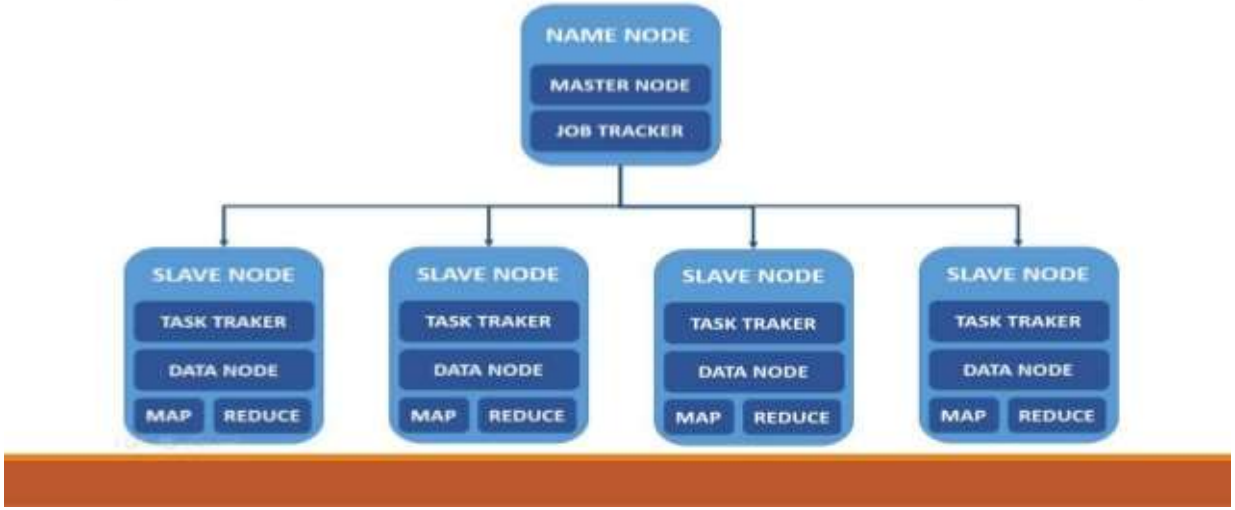
     128MB(New Version)
- Process
  1. MapReduce(Old Model)
  2. Spark(New Model)

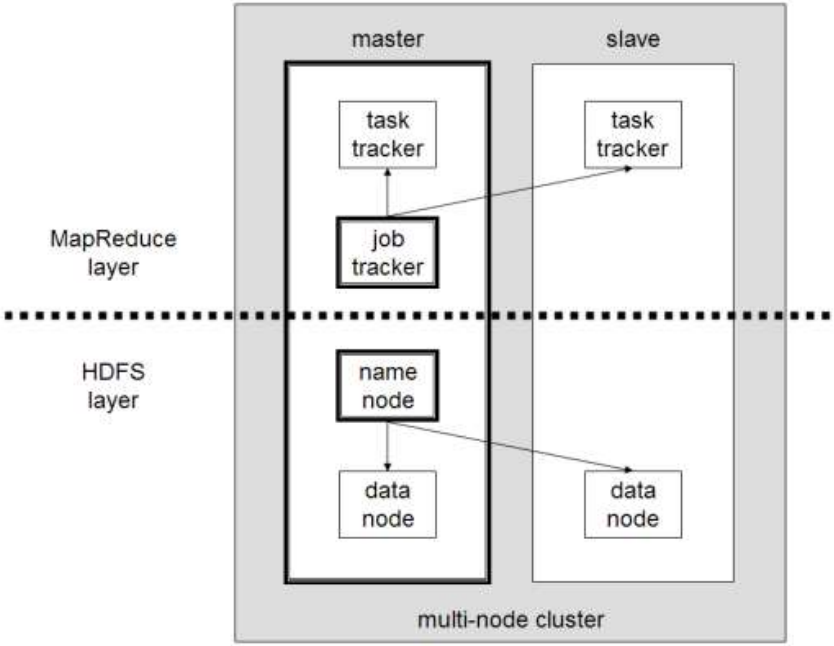
Hadoop Architecture

Hadoop consists of the Hadoop Common package, which provides file system and OS level abstractions, a

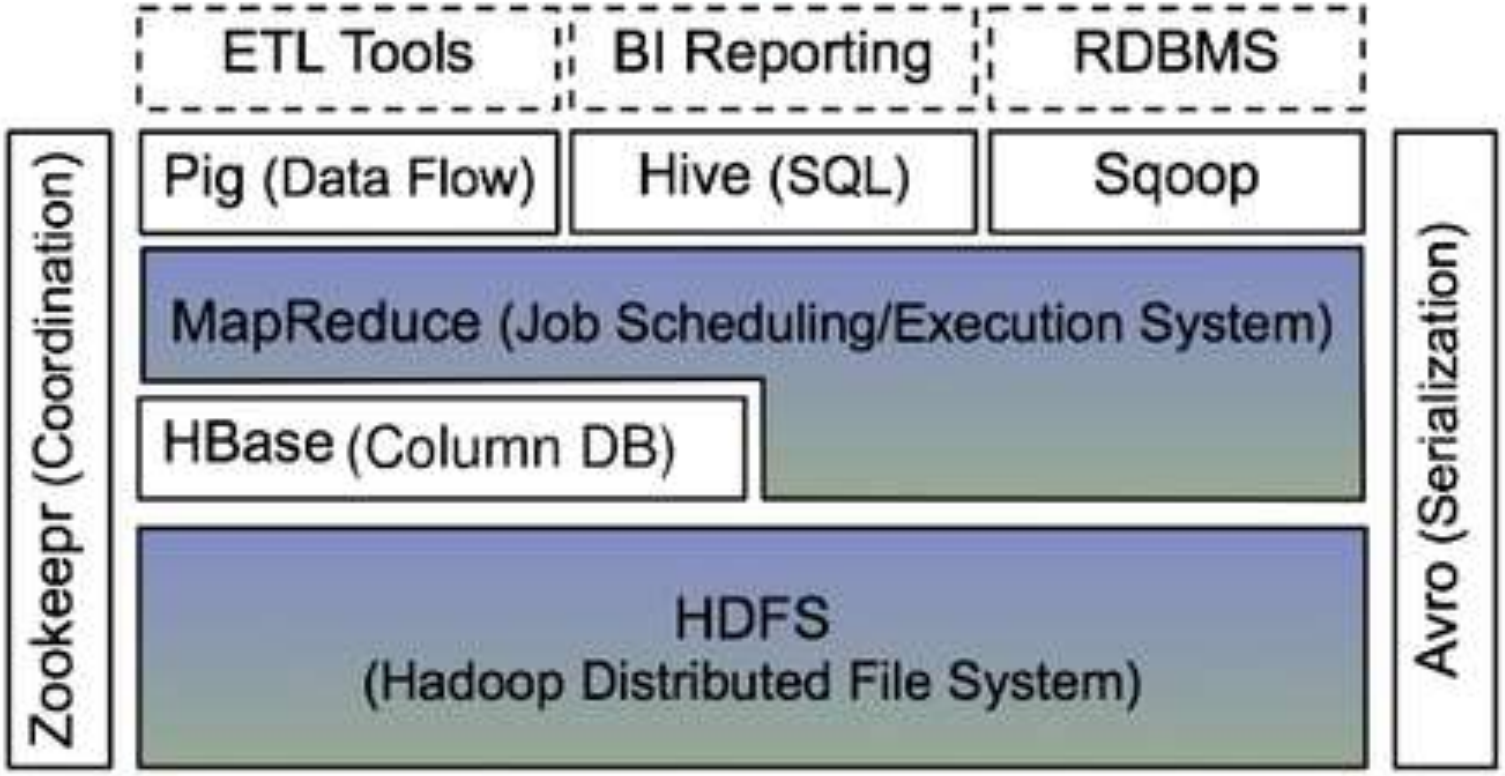
MapReduce engine and the Hadoop Distributed File System (HDFS). The Hadoop Common package contains the necessary Java Archive (JAR) files and scripts needed to start Hadoop.

HADOOP MASTER/SLAVE ARCHITECTURE



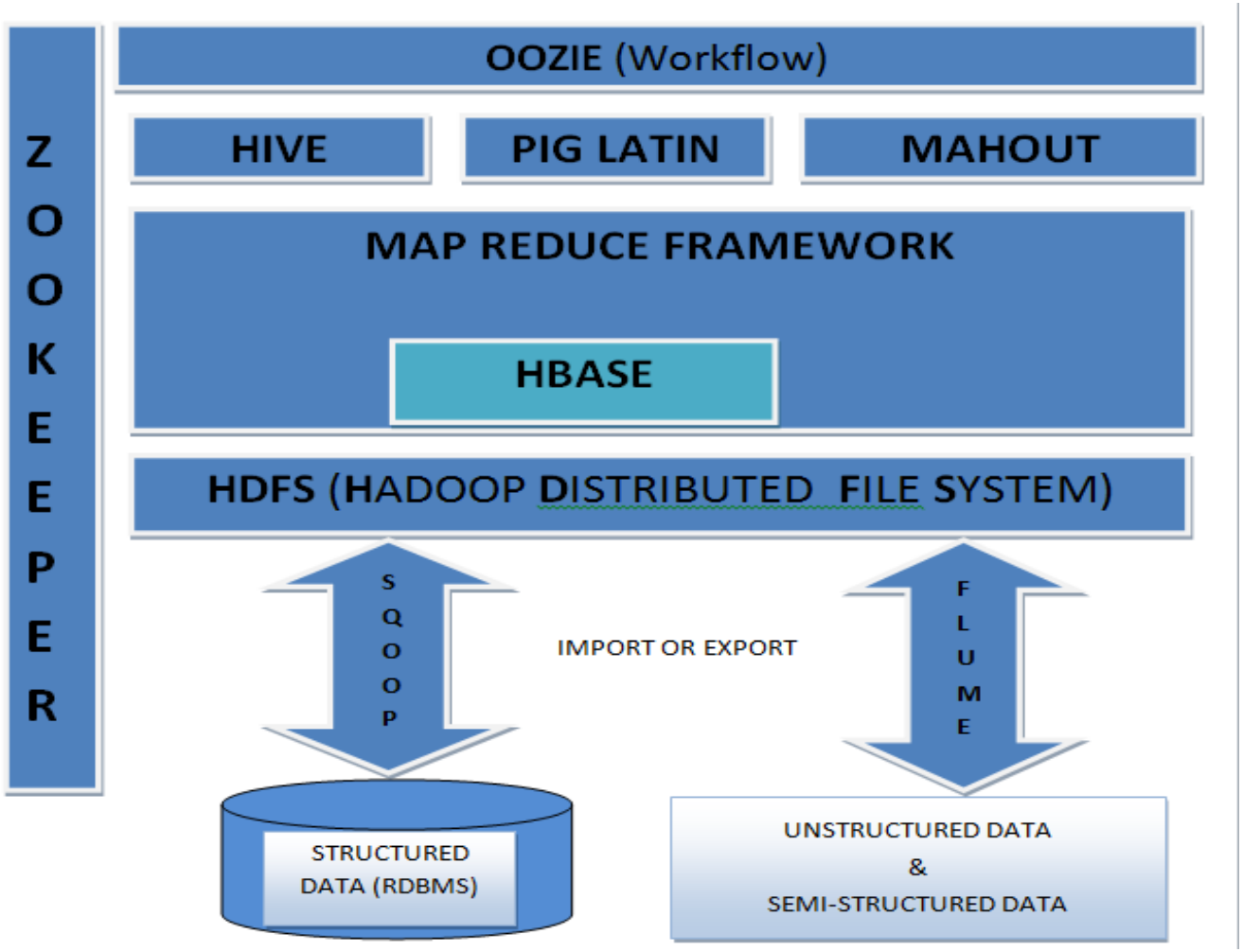


The Hadoop Ecosystem



cloudera

Working Of Ecosystem



HDFS

- Hadoop Distributed File System (HDFS) is designed to reliably store very large files across machines in a large cluster. It is inspired by the GoogleFileSystem.
- Distribute large data file into blocks
- Blocks are managed by different nodes in the cluster
- Each block is replicated on multiple nodes
- Name node stored metadata information about files and blocks

MAPREDUCE

- **The Mapper:-**
  1. Each block is processed in isolation by a map task called mapper
  2. Map task runs on the node where the block is stored
- **The Reducer:-**
  1. Consolidate result from different mappers
  2. Produce final output

HBASE

- Hadoop database for random read/write access
- **Features of HBASE:-**
  1. Type of NoSql database
  2. Strongly consistent read and write
  3. Automatic sharding
  4. Automatic RegionServer failover
  5. Hadoop/HDFS Integration
  6. HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
  7. HBase supports an easy to use Java API for programmatic access.
  8. HBase also supports Thrift and REST for non-Java front-ends.

HIVE

- SQL-like queries and tables on large datasets
- **Features of HIVE:-**
  1. An sql like interface to Hadoop.
  2. Data warehouse infrastructure built on top of Hadoop
  3. Provide data summarization, query and analysis
  4. Query execution via MapReduce
  5. Hive interpreter convert the query to Map reduce format.
  6. Open source project.
  7. Developed by Facebook
  8. Also used by Netflix, Cnet, Digg, eHarmony etc.

PIG

- Data flow language and compiler
- **Features of pig:-**
  1. A scripting platform for processing and analyzing large data sets
  2. Apache Pig allows to write complex MapReduce programs using a simple scripting language.
  3. High level language: Pig Latin
  4. Pig Latin is data flow language.
  5. Pig translate Pig Latin script into MapReduce to execute within Hadoop.
  6. Open source project
  7. Developed by Yahoo

ZOOKEEPER

- Coordination service for distributed applications
- **Features of Zookeeper:-**
  1. Because coordinating distributed systems is a Zoo.
  2. ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

FLUME

- Configurable streaming data collection
- Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

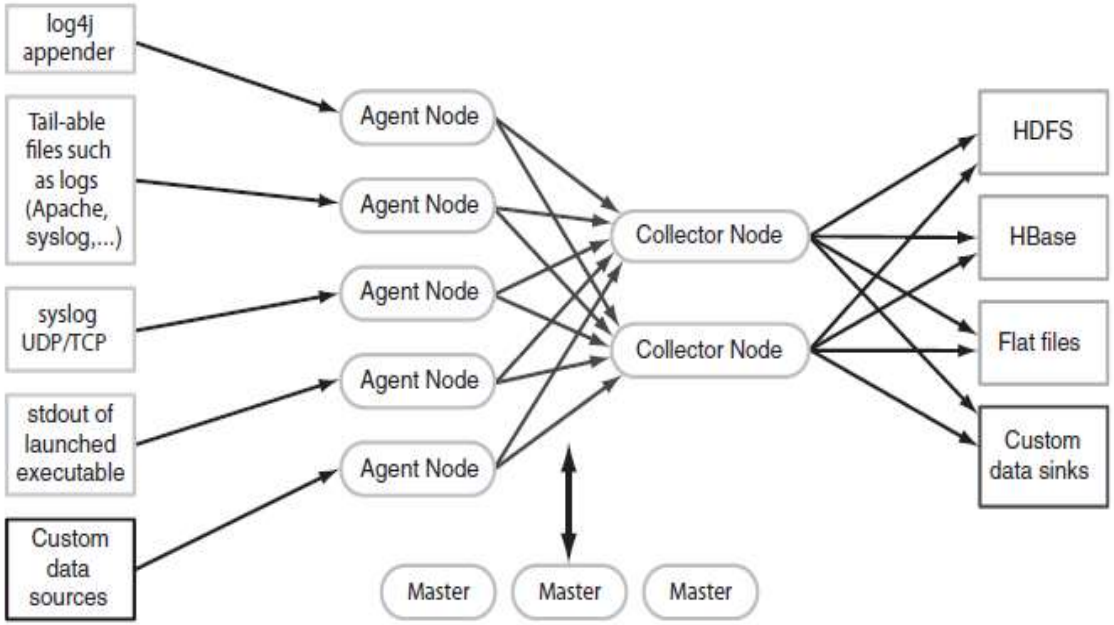
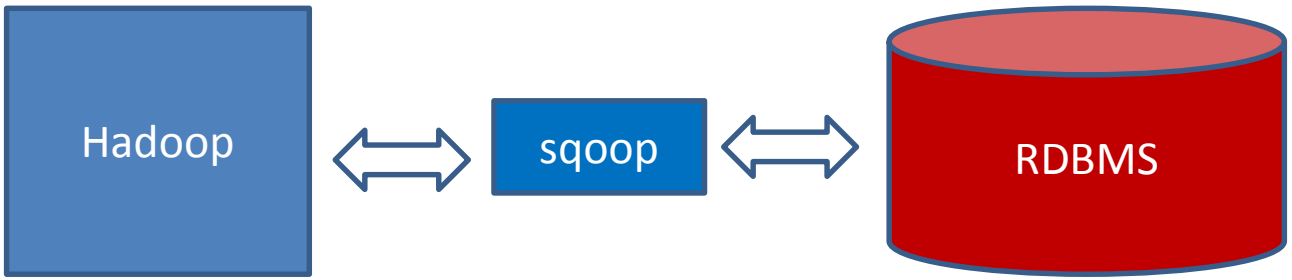


Figure 2.2 **Flume architecture for collecting streaming data**

SQOOP

- Integration of databases and data warehouses with Hadoop
- **Features of Sqoop:-**
  1. Command-line interface for transforming data between relational database and Hadoop
  2. Support incremental imports
  3. Imports use to populate tables in Hadoop
  4. Exports use to put data from Hadoop into relational database such as SQL server



OOZIE

- To design and schedule workflows
- Oozie is a workflow scheduler where the workflows are expressed as Directed Acyclic Graphs. Oozie runs in a Java servlet container Tomcat and makes use of a database to store all the running workflow instances, their states ad variables along with the workflow definitions to manage Hadoop jobs (MapReduce, Sqoop, Pig and Hive).The workflows in Oozie are executed based on data and time dependencies.

Hadoop Advantages

- Unlimited data storage
  1. Server Scaling Mode
     a) Vertical Scale
     b)Horizontal Scale
- High speed processing system
- All varities of data processing
  1. Structural
  2. Unstructural
  3. semi-structural

Disadvantage of Hadoop

- If volume is small then speed of hadoop is bad
- Limitation of hadoop data storage

  Well there is obviously a practical limit. But physically HDFS Block IDs are Java longs so they have a max of 2^63 and if your block size is 64 MB then the maximum size is 512 yottabytes.

- Hadoop should be used for only batch processing
  1. Batch process:-background process

     where user can't interactive
- Hadoop is not used for OLTP
  - OLTP process:-interactive with uses

Conclusion

A scalable fault-tolerant distributed system hadoop for data storage and processing huge amount of data with great speed and maintainence

References

- http://training.cloudera.com/essentials.pdf
- http://en.wikipedia.org/wiki/Apache_Hadoop
- http://practicalanalytics.wordpress.com/2011/11/06/explaining-hadoop-to-management-whats-the-big-data-deal/
- https://developer.yahoo.com/hadoop/tutorial/module1.html
- http://hadoop.apache.org/
- http://wiki.apache.org/hadoop/FrontPage