The Big Data Ecosystem

Tushar B. Kute,
http://tusharkute.com

Hadoop Ecosystem

- Apache Hadoop is an open source framework intended to make interaction with big data easier, However, for those who are not acquainted with this technology, one question arises that what is big data ?
- Big data is a term given to the data sets which can’t be processed in an efficient manner with the help of traditional methodology such as RDBMS.
- Hadoop has made its place in the industries and companies that need to work on large data sets which are sensitive and needs efficient handling.

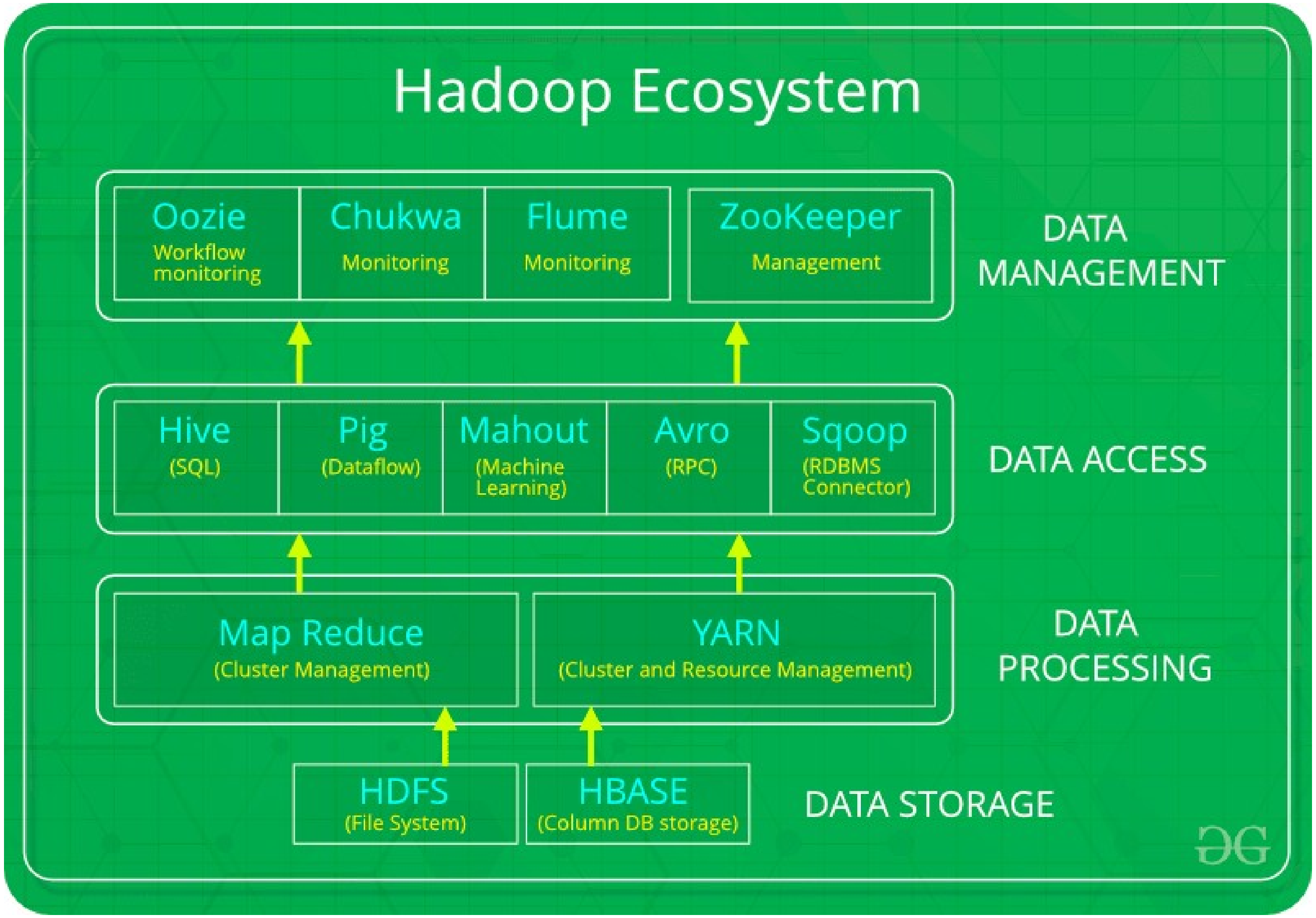
Hadoop Ecosystem

- Hadoop is a framework that enables processing of large data sets which reside in the form of clusters.
- Being a framework, Hadoop is made up of several modules that are supported by a large ecosystem of technologies.
- 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.

Hadoop Ecosystem

- Following are the components that collectively form a 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

Hadoop Ecosystem



Apache Pig

- Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.
- It is a platform for structuring the data flow, processing and analyzing huge data sets.
- Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.
- Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM.
- Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

Hive

- With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).
- It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.
- Similar to the Query Processing frameworks, HIVE too comes with two components: JDBC Drivers and HIVE Command Line.
- JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

Apache Mahout

- Mahout, allows Machine Learnability to a system or application.
- Machine Learning, as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
- It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning.
- It allows invoking algorithms as per our need with the help of its own libraries.

Apache Spark

- It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.
- It consumes in memory resources hence, thus being faster than the prior in terms of optimization.
- Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.

Apache Hbase

- It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database.
- It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.
- At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time.
- At such times, HBase comes handy as it gives us a tolerant way of storing limited data

Solr, Lucene

- These are the two services that perform the task of searching and indexing with the help of some java libraries, especially Lucene is based on Java which allows spell check mechanism, as well. However, Lucene is driven by Solr.

Zookeeper

- There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency, often.
- Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.

Oozie

- Oozie simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit.
- There is two kinds of jobs .i.e Oozie workflow and Oozie coordinator jobs.
- Oozie workflow is the jobs that need to be executed in a sequentially ordered manner whereas Oozie Coordinator jobs are those that are triggered when some data or external stimulus is given to it.

Apache Airflow

- Apache Airflow is a robust platform that allows users to automate tasks with the help of scripts.
- It makes use of a scheduler that helps execute numerous jobs with the help of an array of workers while following a set of specified dependencies.
- Apache Airflow houses rich command-line utilities that allow users to work with DAGs, that help companies order and manage their tasks with ease.

Apache Kafka

- Apache Kafka is one of the most popular open-source software that provides users with a framework to store, read, and analyze streaming data.
- Being open-source, it is available free of cost to users and, hence it houses a broad network of developers & users that help contribute to new features, updates, support functionalities, etc.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com