SE ZG522: Big Data Systems

09 Hadoop Ecosystem

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

April 09, 2022 ONLINE (WILP @ BITS-Pilani Jan-May 2022)

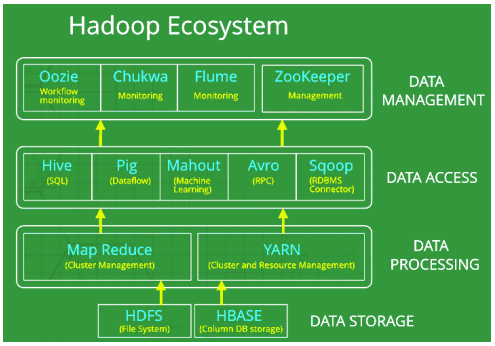
http://ktiwari.in/bds

Hadoop

Hadoop is a framework that enables processing of large data sets residing in the form of clusters.

- Made up of several modules that are supported by a large ecosystem of technologies
  1. **HDFS:** Hadoop Distributed File System
  2. **YARN:** Yet Another Resource Negotiator
  3. **MapReduce:** Programming based Data Processing
  4. **Spark:** In-Memory data processing
  5. **PIG, HIVE:** Query based processing of data services
  6. **HBase:** NoSQL Database
  7. **Mahout, Spark MLLib:** Machine Learning algorithm libraries
  8. **Solar, Lucene:** Searching and Indexing
  9. **Zookeeper:** Managing cluster
  10. **Oozie:** Job Scheduling

Ecosystem



- Task and job tracker, data and name node. 128 MB size block

Components

- **HDFS:** responsible for storing data across various nodes. *Name Node* is the prime node containing metadata (data about data). Data nodes stores the actual data.
- **YARN:** Yet Another Resource Negotiator, is the one who helps to manage the resources across the clusters. Performs scheduling and resource allocation. Three major components 1) Resource Manager, 2) Nodes Manager, 3) Application Manager.
- **MapReduce:** makes the use of two functions i.e. Map() and Reduce()
- **Flume:** Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. Architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.
- **Sqoop:** tool to transfer data between Hadoop and relational database servers (MySQL, Oracle)

Components

- **PIG:** Query language like SQL for structuring data flow, processing and analyzing. Pig Latin language that runs on Pig Runtime. Ease of programming and optimization. Extract Transfer and Load (ETL)
- **HIVE:** highly scalable query language called as HQL (Hive Query Language).
- **Apache Spark:** in memory computations. Does tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc. Suited for real-time data.
- **Apache HBase:** NoSQL database. Provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively. Search of something small is eased. 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.

Components

- **Zookeeper:** Handles issue of management, coordination and synchronization among the resources or the components of Hadoop. Prevents inconsistency by synchronization, inter-component communication, grouping, and maintenance.
- **Oozie:** Oozie performs the task of a scheduler. It binds jobs 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.
- **Kafka:** Distributed streaming platform written in scala.
- **Storm:** Processing engine for real time streaming
- **Ranger:** Framework for data securuty
- **KNOX:** Proxying, Authentication and Client services

Components

- **Mahout:** brings ML to the system. Provides libraries or functionalities such as collaborative filtering, clustering, and classification.
- **Mllib:** scalable ML library consisting of classification, regression, clustering, collaborative filtering, dimensionality reduction, and underlying optimization primitives.
- **Chukwa:** data collection system for monitoring large distributed systems. Includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.
- **Avro:** language-neutral data serialization system to transfer data over a network. A schema-based system.

NoSQL database

- Non SQL, or not only SQL
- Databases that store data in a format other than relational tables
- Key-value stores, Column-oriented databases, Document databases, Graph databases, Multi-Model Databases
- MongoDB, CouchDB, CouchBase, , Amazon SimpleDB, Riak, Lotus Notes are Document-oriented
- Tokyo Cabinet/Tyrant, Redis, Riak, Voldemort, Oracle BDB, Amazon SimpleDB are Key-value stores Databases.
- Cassandra, HBase and Hypertable are Column family stores Databases
- Neo4J, InfoGrid, Infinite Graph, OrientDB, FlockDB are Graph databases

Thank You!

Thank you very much for your attention!

Queries ?