Big Data Tools 1003 (2021F-T1)

Assignment 1

Submitted by: Aadarsha Chapagain

Assignment 1: Provide 3-page summary together on Hadoop, Hive, Sqoop, and PIG.

Hadoop

Hadoop is an open-source framework developed by Apache written in java. It allows the distributed processing of large datasets leveraging the computational power of the commodity hardware. It provides the disturbed computational and storage environment. Scalability is the major strength of Hadoop; it can easily scale from one to thousands of clusters in which each cluster offers storage and computational power.

The major components of Hadoop are

1. Hadoop Distributed File system (HDFS)
2. Yet Another Resource Navigator
3. Map Reduce

HDFS

As its name suggest, it is a distributed file system responsible for storing large datasets which may be either structured or unstructured data. The data are stored among various cluster and the meta data of these nodes or cluster are maintained in log files. HDFS consist of two components

1. Name Node
2. Data Node

Name Node is the major node which store meta data that is data about data. Name node has less resources compared to Data Node as Data Nodes are the ones responsible for the storage and computation. Data nodes are the commodity hardware which makes the Hadoop cost effective.

YARN

Yarn is the resource navigator which manages the resources among the cluster. It allocates and schedules the job in the Hadoop. The three significant components in YARN are

1. Resource Manager
2. Node Manager
3. Application Manager

Resources manager works on allocating the resources to the application and Node Manager allocates the resources such as CPU, bandwidth, memory to the machines. Application Manager acts as a bridge between Node Manager and Resource Manager.

MapReduce

Map reduces uses the distributed and parallel algorithm which changes the large datasets key value pair and perform aggregation on it. It includes two phase that is Map and reduce.

In map phase, sorting and filtering of the data is carried out and a key-value pair is generated as a output which is used as input by reduce phase.

In reduce phase the key value pair from the map phase is taken as input and aggregated to generate the result.

Hive

Hive is data warehouse tool which is distributed and fault tolerant and helps to perform analytics on large datasets. It works on top of Hadoop and makes the analytics and query easy. Hive implements the SQL methodology and interface to perform reading and writing the large datasets. The query language used by hive is called Hive query language (HQL).

Hive supports all the data types that are supported by SQL which makes processing of query easier. Hive can be used for both real-time and batch processing and is scalable. The two major components of Hive are JDBC connectors and Command line tool.

JDBC connectors works on establishing the connection and permission to data storage whereas the command line tool helps in query processing.

Some of the important features of Hive are

- Database and table are built before loading the data
- Hive is used to manage data that are structured and data that resides in table.
- To improve the performance in certain cases, Hive can partition data into directory structures.
- TEXTFILE, SEQUENCEFILE, ORC, RCFILE are some of compatible file formats for hive.
- Hive is built for online Analytical Processing (OLAP)
- Hive has built in User defined function to manipulate strings, dates along with other data-mining tools.

Sqoop

Sqoop is tool used to transfer data between relational databases and Hadoop. It is mainly used to import data from relational databases such as Oracle, MySQL to Hadoop and export data to these databases from Hadoop. Sqoop provides fault tolerance and parallel mechanism. The major function carried out on Sqoop is Sqoop import and Sqoop Export.

Individual tables are imported from RDBMS to HDFS. The row in the table are converted to record in HDFS and the data are stored as text in text files or binary data in sequence or Avro files.

Sqoop export is used to transfer data from HDFS records to RDBMS tables. The records from the set of files are parsed and delimited according to user specified parameter to form row in the table in RDBMS.

PIG

Apache pig is tool used to process the large sets. It provides the high-level abstraction over the map reduce processing. Pig Latin is the scripting language provide by pig to perform the analysis on the datasets. The two major components of the pig tools are

1. Pig Latin
2. Pig Execution Engine

Pig Latin is the scripting language, programmers write the script in this language and execute the script.

Pig Execution Engine provides the abstraction on the map reduce jobs. All the pig scripts are converted to map reduce jobs by execution engine and the results are stored in HDFS.

There are different types of execution environment of pig. Pig local runs on single JVM it is suitable for small data set and distributed environment is used for Hadoop cluster. Pig scripts can be run either on grunt shell or pig server.

Some important features of pig are

- Rich in operators such as join and sort.
- Can process any type of data such as structured, semi structured, and unstructured data.
- Since it is like SQL it is easier to program in PIG.
- Provides higher level of abstraction to Map reduce jobs so that programmers do not have to specify map and reduce jobs.
- Supports full schema, partial schema, and No schema.
- It works on lazy load concept