International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

Big Data and its Tools

Heena Prolta
UG Student, Department of Computer Science and Engineering, UIET, Panjab University, Chandigarh, India

***Abstract*: This research paper is all about knowing the concept of Big Data and Its Tools. Big data includes various tools such as hadoop, sqoop, hive, presto, etc. Along with the introduction of big data all the other concepts which make BIG DATA more useful are included in this paper. This paper include characteristics of big data as well as some key challenges to big data. The two tools of big data which are used on large scale are Hive and Sqoop. Introduction of hive and sqoop with all the commands are very well explained in this paper.**

***Keywords*: Validating data, big data and tools**

1. Introduction

We all are very familiar with the fact that all the companies uses data since long and uses various tools to store the data. The quantity of data on this planet is growing exponentially for many reasons.

What is big data?

BIG DATA is a term which is used for various collection of data sets which are very large and complex and they are very difficult to store and process using the available database management tools.

Characteristics of big data

Basically, there are three characteristics of big data:
Volume
Variety
Velocity

Volume

It defines the amount of data which is exponentially increasing day by day at a very fast rate. Data can be generated by anything such as machines, humans and the interactions on social media. According to research, around 40 zettabytes will be generated up to 2020. This largeness of data is itself a challenge for storing.

Variety

Data can be structured as well unstructured. Data is available in different forms such as texts, images, videos, tables. Etc. Earlier, most of the data is generated by excel and database. But, today data is generated by audios, videos, images, sensors etc.

Velocity

The speed at which data is generated every day is termed as velocity of data. Data is very massive and flows at a very fast rate. If we talk about Facebook, then there 1.03 billion active users of Facebook per day. This is almost 22% increase year by year.

Types of big data

Structured, unstructured and semi-structured.

Structured

The data which can be stored and processed in some fixed format is known as structured data. This data is stored in RDBMS (relational database management system), SQL (structured query language), etc.

Unstructured

Data which is of the unknown form and cannot be stored in RDBMS is known as unstructured data. This data cannot be analyzed unless and until it is transformed into structured data.

Semi structured

It includes that type of data which do not have a formal structure of data model like DBMS. XML files, JSON documents are some examples of semi-structured data.

2. Challenges in big data

A. Security

We know that big data means large amount of data. To store this large data in a secure form is itself a challenge. it includes various factors such as authentication, restricted access for some users

B. Data storage

We all know that volume of data is arrive in that in order to store it. To solve this directly proportional to complexity. The more the data is, the more problems storage problem we need a system to store it.

C. Lack of talent

We have lots of big data projects in the organizations. It is still a challenge in the field that we still don't have very well data scientists, developers and analysts.

D. Validating data

Data integration's concept is used in data validation. We noticed that many organizations get similar data in small pieces from different systems and that data from different systems doesn't always agree. For example, the ecommerce system may show daily sales at a certain level while the enterprise resource planning (ERP) system has a slightly different number.

3. Tools of big data

Hadoop

Hadoop is an Apache open source framework written in java which allows distributed processing of large datasets across the clusters of computers using simple programming models. Hadoop application works in that environment which provides distributed storage and computation across the clusters of computers.

Hadoop distributed file system (HDFS)

HDFS is based on the Google file system (GFS) which provides a distributed file system to run on a commodity hardware. It has more similarities as compared to existing distributed file systems. There are various differences in HDFS as compared to simple distributed file system. They are very fault- tolerant and are designed to be deployed on low cost software.

3.1 Hive

Hive is a data ware house infrastructure tool which is used to process structure data in hadoop. Earlier, Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used store Schema in a database and processed data into HDFS. It provides SQL type SQL type language for querying called HiveQL or HQL.

HiveQL:Database Queries

-Show databases;
-Create <external> table <table-name> (eid int, name string, salary string, designation string)
>Row format delimited
>Fields terminated by '\t'
>Lines terminated by '\n'
>stored as textfile;
-drop table <table-name>;

Insert Data into Table

- **Load data from LFS**

-load data local inpath <'FilePath>' into table<table-name>;

- **Load data from LFS**

-load data inpath <'FilePath>' into table<table-name>;

Hive Script

-vi createEx.hql
use <database name>;
create table emp(eid int, ename string, esal double)
>row format delimited
>fields terminated by '\t'
>lines terminated by '\n'
>stored as text file;
-Execute Hive Script:
hive –f "createEx.hql"

3.2 SQOOP

Sqoop is a tool which was designed to transfer data between Hadoop and relational database servers. It import data from relational database to HDFS and export data from HDFS to relational database.

SQOOP import

The sqoop import tool is used to import individual tables from RDBMS to HDFS. Each row in the table is treated as a record in HDFS. All the records are stored as text data in text files.

SQOOP export

The sqoop export tool is used to export the set of files from HDFS to RDBMS. The files is given as input to sqoop containing records, which are called as rows in a table.

Basic operation for import:
RDBMS to HDFS:

mysql –u root –p
Enter password: cloudera mysql> show databases;
mysql>create database sqoopDB; mysql>use sqoopDB;
mysql>show tables;
mysql>create table emp(eid int(5), ename varchar(20), salary int(100);
mysql>desc emp;
mysql>insert into emp values(101, 'abc', 10000); mysql>select *from emp;

Copy RDBMS data in HDFS location

mysql>grant all privileges on sqoopDB * to 'Qlocalhost';
mysql>sqoopimport –connect jdbc:
mysql://localhost/sqoopDB—table emp –m 1-targetdire/xyz
hadoop fs –ls /xyz
Sqoop import –connect jdbc:
mysql://localhost/sqoopDB—table emp –m targetdir /xyz1 –fields-terminated-by-'\t';
Sqoop import –connect jdbc: mysql://localhost/sqoopDB—table emp –m 1— targetdir /xyz2 –fields-terminated by'\t' –columns 'eid.name' –where 'salary>1200;

Basic Operations for export:
HDFS to RDBMS:

Vi emp.txt
110,asdf,1500
111,qwer,2100
112,sdfg,3400
hadoop fs –put emp.txt /xyz
sqoop export –connect jdbc: mysql://localhost/ sqoopDB table emp –export-dir /xyz/emp.txt

4. Conclusion

From this paper, I concluded that big data is a large amount of data and it faces different challenges regarding to storage, security, validating, etc. Today, all the industries, hotels, or any other field deals with big data. Production of big data is

increasing at a very massive rate and with a very high velocity. This paper mainly focuses on the tools of big data i.e. HIVE and SQOOP. Hive is a data ware house infrastructure tool which is used to process structure data in hadoop. Sqoop is a tool which was designed to transfer data between Hadoop and relational database servers. The programming concept of these tools is explained with the required commands.

References

[1] M. A. Beyer and D. Laney, "The importance of "big data": A definition,"
Gartner, Tech. Rep., 2012.
[2] https://dzone.com/articles/sqoop-import-data- from-mysql-to-hive
[3] https://www.edureka.co/blog/big-data-tutorial
[4] https://www.datamation.com/big-data/big-data-challenges.html
[5] https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm