Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230798

Understand the working of Sqoop and hive in Hadoop

Archana Uriti, Surya Prakash Yalla, Chunduru Anilkumar

Department of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh, India

archana.u@gmrit.edu.in

Abstract. In past decades, the structured and consistent data analysis has seen huge success. It is a challenging task to analyse the multimedia data which is in unstructured format. Here the big data defines the huge volume of data that can be processed in distributed format. The big data can be analysed by using the hadoop tool which contains the Hadoop Distributed File System (HDFS) storage space and inbuilt several components are there. Hadoop manages the distributed data which is placed in the form of cluster analysis of data itself. In this, it shows the working of Sqoop and Hive in hadoop. Sqoop (SQL-to-Hadoop) is one of the Hadoop component that is designed to efficiently imports the huge data from traditional database to HDFS and vice versa. Hive is an open source software for managing large data files that is stored in HDFS. To show the working, here we are taking the application Instagram which is a most popular social media. In this analyze the data that is generated from Instagram that can be mined and utilized by using Sqoop and Hive. By this, prove that sqoop and hive can give results efficiently. This paper gives the details of sqoop and hive working in hadoop.

Keywords: big data, Hadoop, Sqoop, Hive, HDFS.

1. Introduction

Generally huge data cannot be handled with the traditional database system tool. Big data contains any kind of data like structured, unstructured and semi structured. To manage the huge databases in terms of analysis, storage, sharing, searching, analysis and visualization, the parallel softwares are required. Firstly, consider the data from various sources like social media, business, etc. Flume is one of the component in hadoop and is used to acquire data from social media such as twitter. Then, this data can be organized using distributed file systems such as Google File System or Hadoop File System which are efficient when more number of reads are there compared to writes. Finally, with the help of map reduce the data can be analysed and queries can be run efficiently shown in figure 1.

DOI: 10.54254/2755-2721/6/20230798

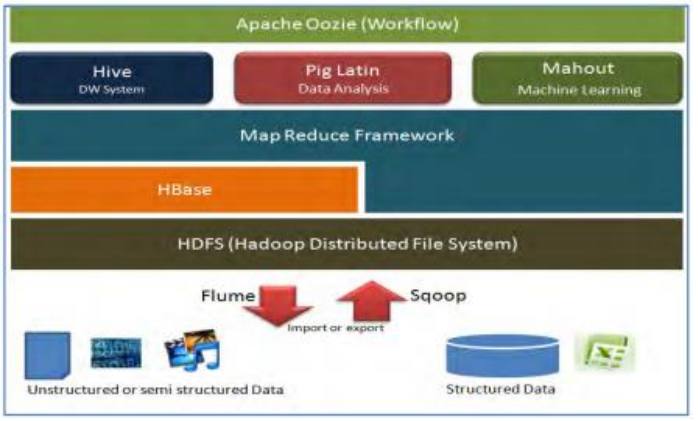

Figure 1. Hadoop ecosystem.

The most advanced analytic technique is big data analytics which is used to analyse large data sets that contains structured, unstructured and semi-structured format of data. This technique helps the various business people to take faster decisions and to increase their profits. Hadoop is the best tool to analyse huge data and it is a distributed parallel process. The hadoop tool contains four major components that are hdfs for data storage, yarn for resource allocation, map reduce for data processing and other is hadoop common. Along with these four components, other components can be included that are commonly called as ecosystem. In this, considered the fast accessing hadoop component that is Apache Hive and other component is sqoop which is used to transfer the data from rdms to hdfs. These two components are described in methodology. This paper mainly aims to efficient access of data with the help of sqoop and hive.

2. Literature survey

Sidhu, R et al. (2015) shows the working of hadoop by installing it and also implemented the word count example using mapreduce which will impact the runtime of the program. It focus in the gaining the knowledge on hadoop and sqoop. Here considering the limited dataset to copy into the hdfs and test the program in the given time. It shows the importance of sqoop among all other components available [1]. Daoqu Geng et al. (2019) proposed a platform for industries which helps to decrease the processing time with less data storage space. Here focusing on evaluating the impact of multiple compression and serialization methods on big data platforms. In that consider the optimal compression and serialization methods. The results shown that compression time is reduced by 73.9% with a less than 96% by comparing the methods which is integrated in hadoop and spark. Also it shows the reduction in data serialization by 80.8% [2].Urmila R. Pol (2016) presents a study on analyzing the big data using hadoop tool. It shows that each working of mapreduce, pig and hive. Also shows the importance and drawback of these components. Pig and Hive are having less lines of code when compared to mapreduce. Hive offers less optimization and control on the data flow than pig [3]. Anisha P. Rodrigues et al. (2018) proposed a method to find the recent trends in tweets and perform analysis using pig and hive. It shows the execution time of each result obtained from pig and hive. Finally performance is defined by taking the less execution time from the twitter data analysis. It is concluded that pig is more efficient than hive [4]. Rida Qayyum (2020) describes that big data plays a huge role in handling large amount of data. It defines the various problems involved in big data and then provide the opportunities. To handle the big data, hadoop plays a key role and solves the various issues involved in big data which is taken from different platforms [5].

3. Proposed methodology

The proposed system tries to implement the results with accuracy while performing query on the dataset using Hive commands [6]. Also it focus on the working of sqoop to transfer the data from rdms to hadoop which helps the users to modify the dataset according to their requirements. Here Hive is

DOI: 10.54254/2755-2721/6/20230798

one of the free source component of Hadoop having their own query language called Hive Query Language [7].

It is designed in such a way that it can analyse the data set and can give accurate outputs to the query processed. These Hive commands are simple to use and keeps the query to run fast and also it takes less time to write the query. Here considering the sample dataset from the most running social media Instagram[8]. The below is the figure 2 shows the working of sqoop. In this, this working can be represented with the consideration of dataset [9].

In order to get more efficiency of the Hadoop framework and Sqoop technologies, here hadoop is installed in cloudera. By this, one can easily show and implement the work. The huge data can be accessed efficiently and also gives the more performance [10].

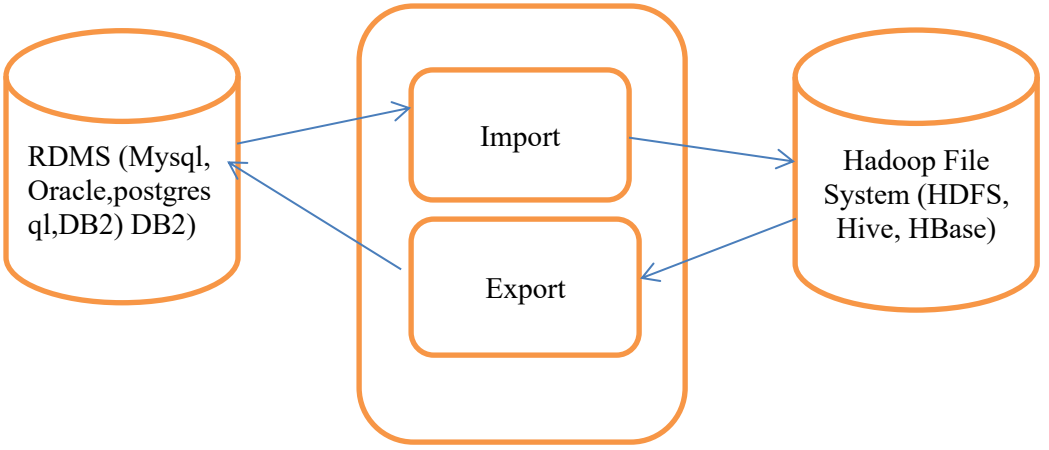

Figure 2. Flow chart of sqoop profile.

When considering Sqoop, it has an advantage of simple data structures to import or export the data from relational data structure to hadoop and vice versa. The Sqoop can easily move data from the RDBMS to the Hadoop file system. Some benefit of sqoop is using simple commands for transferring the data that can be implemented from command line interface, allows the client side installation, for data preprocessing take the help of hive and hdfs. Sqoop has the benefit to transfer the data, but it also has drawback like giving inefficient when the user manually enter the data. The below figure 3 represents the data acquisition scheme of the data platform by sqoop.

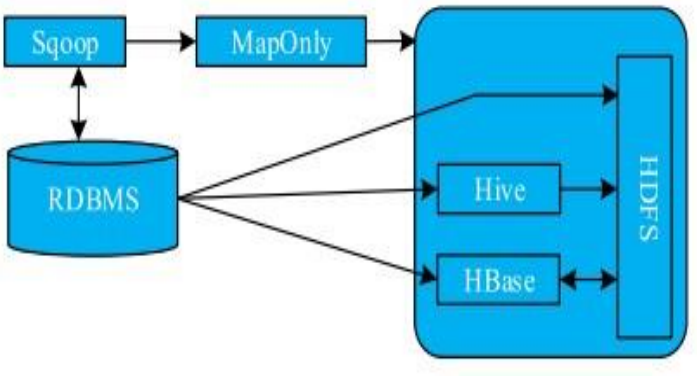

Figure 3. Data Acquisition scheme of the data platform by Sqoop.

4. Implementation

STEP-1: Login into MySQL

To store the data in hdfs, use the Apache Hive acts as sql interface between the user and the Hadoop distributed file system which integrates Hadoop that is shown in below figure 4. Here use the command " mysql –u root training"

DOI: 10.54254/2755-2721/6/20230798

```
[training@localhost ~]$ mysql -u root training
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 4
Server version: 5.0.77 Source distribution

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.
```

Figure 4. Login into MySQL.

STEP-2: After login into mysql, database is created and then create a table with values.

In this step first, create database and then create our own dataset to perform the Hive commands. The following command to create database and the dataset.

create database instagram;

create table instagram.insta(profile_name varchar(65),followers int, following int,posts int);

Creation of database and dataset are shown in below figure 5.

```
mysql> create database Instagram;
Query OK, 1 row affected (0.00 sec)

mysql> create table Instagram.insta(profile_name varchar(65),followers int,following int,posts int);
Query OK, 0 rows affected (0.00 sec)

mysql> insert into Instagram.insta values("rahul",295000,360,56);
Query OK, 1 row affected (0.00 sec)

mysql> insert into Instagram.insta values("kohli",495000,60,42);
Query OK, 1 row affected (0.00 sec)

mysql> insert into Instagram.insta values("rohith",295567,90,142);
Query OK, 1 row affected (0.00 sec)

mysql> insert into Instagram.insta values("surya",25590,142,34);
Query OK, 1 row affected (0.00 sec)

mysql> insert into Instagram.insta values("axar",24560,182,94);
Query OK, 1 row affected (0.00 sec)

mysql> insert into Instagram.insta values("Dhoni",234560,167,18);
Query OK, 1 row affected (0.00 sec)
```

Figure 5. Creation of database and dataset.

STEP-3: Table View

Here to check whether the table has created or not, for this use the following command

select * from instagram.insta;

STEP-4: In hive, need to create a database and a table to store the imported data that is shown in figure 6.Here exit from the mysql and login into the hive in order to import the created table in mysql.

```
hive> create database instagramdata;
OK
Time taken: 1.921 seconds
hive> use instagramdata;
OK
Time taken: 0.013 seconds
hive> create table instagram(profile_name string,followers int,following int,posts int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n';
OK
Time taken: 0.257 seconds
```

Figure 6. Creation of database and table in hive.

STEP-5: Run the import command on Hadoop to transfer the data from RDBMS to HDFS

Sqoop import --connect \ jdbc:mysql://147.0.0.1:3006/instagram \ --username root --password training \ --table Instagram.insta \ --hive-import --hive-table insta_hive.insta_hive_table \ --m 1

where 147.0.0.1 is IP address and 3006 is the mysql port number

DOI: 10.54254/2755-2721/6/20230798

STEP-6 : In hive, check whether the data is transferred successfully or not that is shown in below figure 7. By using the command, check whether the data is imported into hive successfully or not

Select * from instagram;

```
hive> load data local inpath '/home/training/insta.csv' into table instagram;
Copying data from file:/home/training/insta.csv
Copying file: file:/home/training/insta.csv
Loading data to table instagramdata.instagram
OK
Time taken: 0.182 seconds
hive> select * from instagram;
OK
profile_name    NULL    NULL    NULL
rahul   295000  360     56
kohli   495000  60      42
rohith  295567  90      142
surya   25590   142     34
axar    24560   182     94
Dhoni   234560  167     18
shami   34560   67      89
jadeja  35660   77      39
dinesh  345000  460     320
bumrah  455000  45      69
Time taken: 0.142 seconds
```

Figure 7. data imported into hive.

5. Results and analysis

Now working on the dataset with hive commands

hive>select * profile_name, followers from Instagram sort by followers DESC LIMIT 5; shown in figure 8a and 8b.

```
hive> select profile_name,followers from instagram sort by followers DESC limit 5;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202210302316_0001, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202210302316_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202210302316_0001
2022-10-31 00:09:52,986 Stage-1 map = 0%,  reduce = 0%
2022-10-31 00:09:55,000 Stage-1 map = 100%,  reduce = 0%
2022-10-31 00:10:02,035 Stage-1 map = 100%,  reduce = 33%
2022-10-31 00:10:03,044 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202210302316_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202210302316_0002, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202210302316_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202210302316_0002
2022-10-31 00:10:05,305 Stage-2 map = 0%,  reduce = 0%
2022-10-31 00:10:07,964 Stage-2 map = 100%,  reduce = 0%
2022-10-31 00:10:16,002 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_202210302316_0002
OK
```

Figure 8a. profile name, followers taken from table by using hive command.

```
kohli   495000
bumrah  455000
dinesh  345000
rohith  295567
rahul   295000
Time taken: 28.436 seconds
```

Figure 8b. output of profile name, followers.

hive>select * profile_name, followers from Instagram WHERE followers = 59557; shown in figure 9.

DOI: 10.54254/2755-2721/6/20230798

```
hive> select profile_name,followers from instagram where followers=59557;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_202210302316_0008, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202210302316_0008
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202210302316_0008
2022-10-31 00:23:10,024 Stage-1 map = 0%,  reduce = 0%
2022-10-31 00:23:11,027 Stage-1 map = 100%,  reduce = 0%
2022-10-31 00:23:12,032 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202210302316_0008
OK
Time taken: 5.1 seconds
```

Figure 9. profile name, followers extracted where followers limit=59557

6. Conclusion

In this, the main goal is to show the working of sqoop and hive in hadoop by taking the application of Instagram. After importing the data from rdbms to hdfs with the help of sqoop. Analysis of data can be possibly done with the help of Hive. Hive is beneficial for large data sets accessing in a faster manner with the help of hive query language. Fault tolerance is main benefit of Hive and gives the results efficiently. By using Hive commands fetch the results accurately for the given dataset within less time. Finally concluded that sqoop is working effectively in transferring data between rdms and hdfs. Also shows the efficient working of Hive.

References

[1] Sidhu, R., Chea, D., Dhakal, R., Hur, A., & Zhang, M. (2015). Implementation of Hadoop and Sqoop for Big Data Organization.

[2] Geng, D., Zhang, C., Xia, C., Xia, X., Liu, Q., & Fu, X. (2019). Big data-based improved data acquisition and storage system for designing industrial data platform. IEEE Access, 7, 44574-44582..

[3] Pol, U. R. (2016). Big data analysis: comparison of hadoop mapreduce, pig and hive. International Journal of Innovative Research in Science, Engineering and Technology, 5(6), 9687-93.

[4] Rodrigues, A. P., & Chiplunkar, N. N. (2018). Real-time Twitter data analysis using Hadoop ecosystem. Cogent Engineering, 5(1), 1534519.

[5] Qayyum, R. (2020). A roadmap towards big data opportunities, emerging issues and hadoop as a solution. Rida Qayyum." A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution", International Journal of Education and Management Engineering (IJEME), 10(4), 8-17.

[6] Xiao, H., VE, S., & Manickam, A. (2022). Research of College English Online Course Based on Cloud Computing and Exploitation for Multimedia Asian Information Processing. ACM Transactions on Asian and Low-Resource Language Information Processing.

[7] Rajalaxmi, R. R., Saradha, M., Fathima, S. K., Sathish Kumar, V. E., Sandeep kumar, M., & Prabhu, J. (2022). An Improved MangoNet Architecture Using Harris Hawks Optimization for Fruit Classification with Uncertainty Estimation. Journal of Uncertain Systems.

[8] Liu, Y., Sathishkumar, V. E., & Manickam, A. (2022). Augmented reality technology based on school physical education training. Computers and Electrical Engineering, 99, 107807.

[9] Easwaramoorthy, S., Moorthy, U., Kumar, C. A., Bhushan, S. B., & Sadagopan, V. (2017, January). Content based image retrieval with enhanced privacy in cloud using apache spark. In International Conference on Data Science Analytics and Applications (pp. 114-128). Springer, Singapore.

[10] VE, S., Park, J., & Cho, Y. (2020). Seoul bike trip duration prediction using data mining techniques. IET Intelligent Transport Systems, 14(11), 1465-1474.