Transfer of Digital Library Data Among Different Databases by Big Data Tool "Sqoop"

[1]**Reshu,** [2]**Mr. Arun Mittal**

[1]Student M. Tech, Greater Noida Institute of Engineering and Technology, Greater Noida (India)
[2]Asst.Professor, Greater Noida Institute of Engineering and Technology, Greater Noida (India)

Abstract: Digital Libraries are the integral part for storing and accessing huge amount of data. Most of the digital libraries are using the traditional relational database management system (RDBMS) for storing, updating, modifying, deleting and fetching the information such as MySQL, Oracle RDB, SQLite, DB2, etc. But now due to the advancement of technologies like Big Data, it has become important to transfer the data from traditional database management System (RDMS) to HDFS (Hadoop Distributed File System), HIVE or HBase, etc. This transfer of data among different databases can be achieved by a Big Data tool named as “Apache Sqoop”. This tool helps in importing the data from RDBMS to HDFS and exports the data from HDFS to RDBMS as per the requirement.

In this paper we will understand the importance of data transfer among different databases, we will also study about the definition of the Sqoop, its architecture, its feature, its features. We will also see the import and export functionality of the sqoop for transferring the data. In end, we will discuss about the future scope and the limitation of using these tools in the Digital Library.

***Keywords*- Digital Library, Big Data, Sqoop.**

I. Introduction

Libraries are most important department for gathering and accessing all the useful information related to different subjects. Due to digitalization and with continuous development in information technology, it has become important that the library data should be handled in a proper manner as digital libraries has started using advanced technologies Big Data in order to deal with varied and huge data available in the Library.

Tool of Big Data, which is used for transferring the data among Relational Databases Hadoop system is known as Sqoop. Sqoop can import the data from RDMS to Hadoop Distributed File System (HDFS) and can also export the data from HDFS to RDMS. Sqoop used the concept of MapReduce to import and export the data. It provides fault tolerance and parallel operation.

It has below mentioned capabilities as well :-

1) Import individual database table or the entire database to the filed present in HDFS.
2) It also generated the java class in order to interact with the imported data
3) It has the capability to import the data from RDMS to directly into Hive or NOSQL database – Hbase.

We use **import** command to import the data from RDBMS to Hadoop and **export** command to export the data from Hadoop to RDMS.

In order to work with the latest technology like Big Data we should also use the tools for automating our processing of data transfer from external environment (RDMS) to the internal environment (HDFS).

Literature Review

In this section we will be discussing about the definition, features and Sqoop tool in order to use in digital libraries.

2.1 Sqoop – Defination

Big Data tool Apache Sqoop is an open source software [1]. Command line interface is used by Sqoop for providing the efficient transfer to huge amount of data among Apache Hadoop and RDMS.

2.2 Sqoop – Features

Sqoop provides multiple advantages because of its varied features:-

1) Provide Fast Performance
2) Provide Fault Tolerance
3) System utilization in optimal manner
4) Imported data is transferred by MapReduce process.
5) Easy integration with other databases like Pig, Hive, HBase [2].
6) Perform the import and export in parallel
7) Incremental Load is possible i.e some part of the table can be loaded.
8) Full Load is possible that means a full table can be loaded by using a single command of the sqoop.
9) By using gzip algorithm, data can be compressed.
10) By Sqoop data can be imported into Accumulo.

.

II. SQOOP ARCHITECTURE

In this section, we will discuss about the architecture and working of the Sqoop.

Before discussing the architecture of the Sqoop, it is interesting to understand that from where the name has been derived from "SQL to Hadoop & Hadoop to SQL".

Most of the existing database system has been designed with the understanding of standard database management system like SQL but every database has some different feature which makes it different from one another. Because of these different features, it becomes difficult to transfer the data across different databases. Big Data tool Sqoop make this job easier. [3]

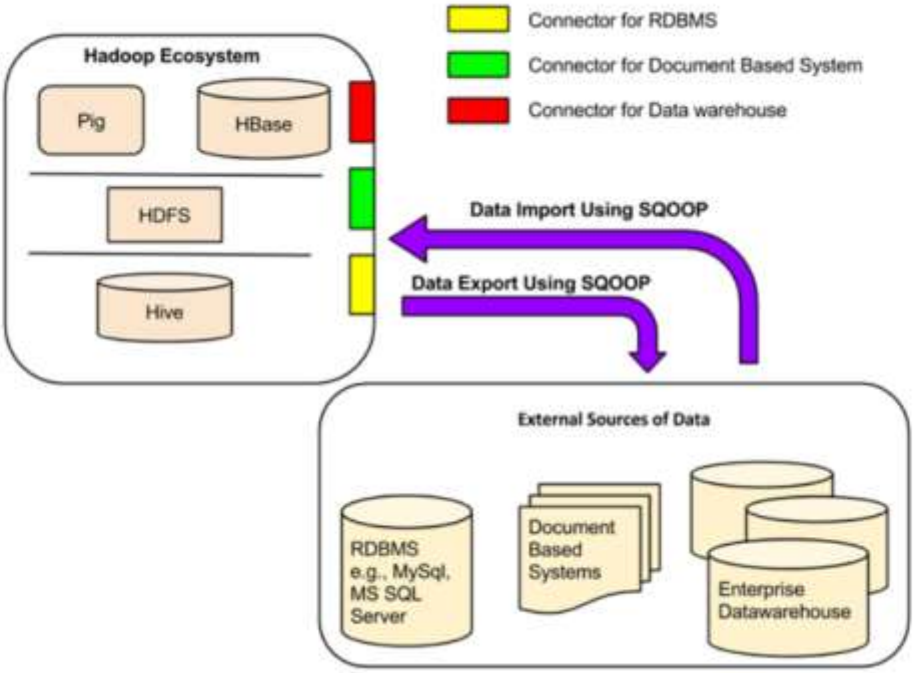


Fig 1 - Sqoop Architecture

Sqoop connectors are used for data transfer among different databases easily like from SQL to HDFS, Hive, Pig etc. Sqoop has connector which can work with many famous relational databases; there is one JDBC connector also which connect to the database which supports JDBC's protocols. It also consists of many third party connectors for data storage. In Fig 2, we can see the working of the sqoop, the main task is divided into multiple subtask when the sqoop command is submitted. These map task handled the work individually internally, the subtask is the one which import the data to the Hadoop Ecosystem. It means that map task is the one which import the full data.

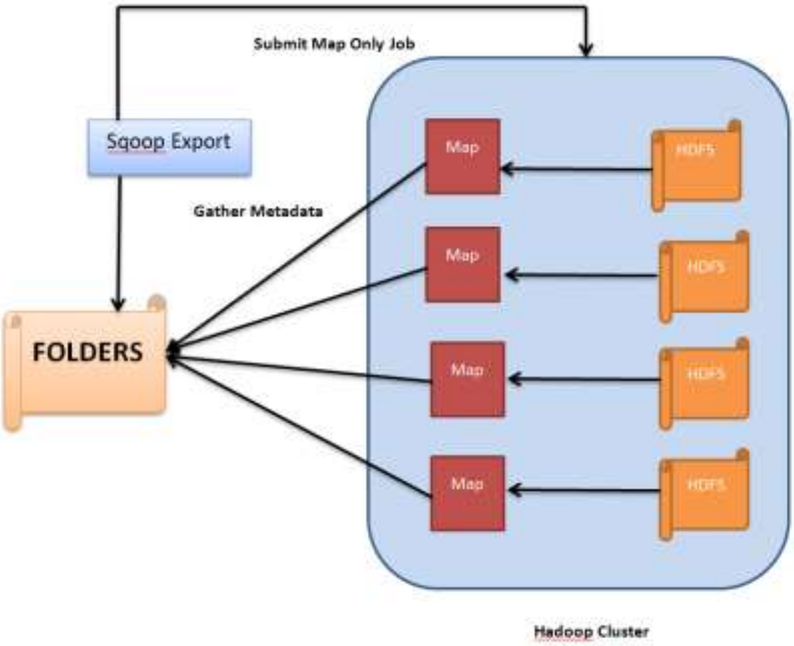


Fig 2 – Sqoop Working

Tool which is used to export the set of file from HDFS to back to RDMS is called sqoop export tool. [4]
The job is mapped into the map task when we submit the job, it bring the data in chunks into the HDFS and then these chunk data is exported to a structured data at the desired destination. In this way, we can get all the data by combining all the data which are present in the chunks.
Aggregation of the data is not done by the sqoop, it is only responsible for importing and exporting the data.

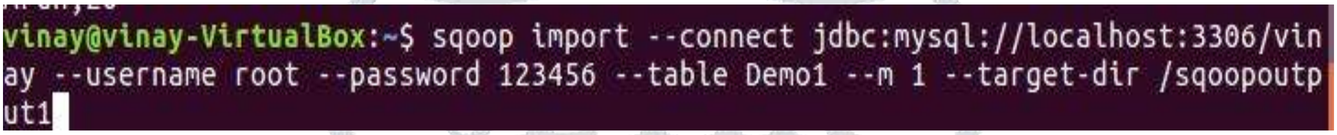
III. SQOOP – TRANSFER OF DATA AMONG DATABASES

In this section we will learn about import and export of the data among different databases. We will discuss about following transfers:-

4.1 Sqoop - Transfer of traditional data of RDBM to Hadoop

In order to transfer the data from traditional relational database management system to hadoop, a tool is required which is known as sqoop. It helps in importing data from RDMS, MySQL, Oracle or Mainframe to HDFS. It helps in automation of the data transfer in a faster way because it use MapReduce to import or export data in parallel manner.
Some of the Steps of the Sqooping the data:

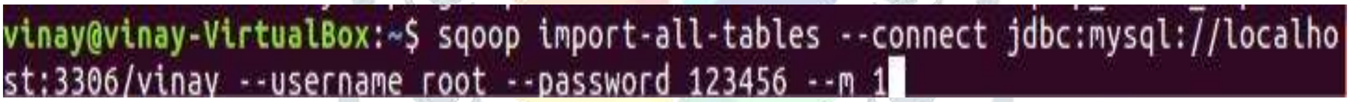
1) Import table from mysql to hdfs



```
vinay@vinay-VirtualBox:~$ sqoop import --connect jdbc:mysql://localhost:3306/vin
ay --username root --password 123456 --table Demo1 --m 1 --target-dir /sqoopoutp
ut1
```

In above screen shot table Demo1 had been created under vinay database which was imported to hdfs i.e sqoopoutput1 directory

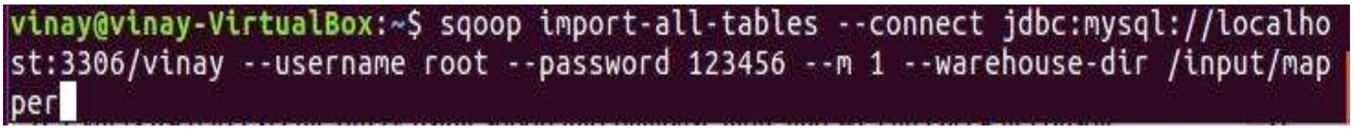
2) Import all tables from mysql to hdfs



```
vinay@vinay-VirtualBox:~$ sqoop import-all-tables --connect jdbc:mysql://localho
st:3306/vinay --username root --password 123456 --m 1
```

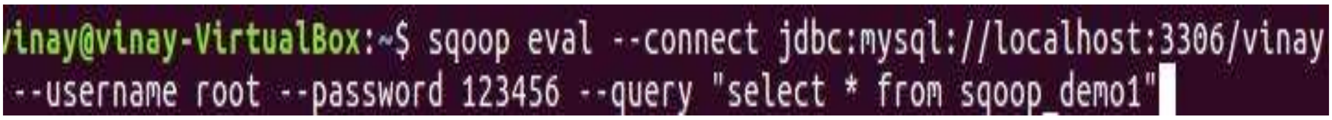
All tables under vinay database imported to hdfs . In above command we have not specified target directory because it does not work with import-all-tables. It imports all tables in default path of hdfs. We can use –warehouse-dir argument to import all tables in specified directory.

Note: If you are using import-all-tables .All tables should have a primary key
By default 4 –m i.e mappers get created which are used to move data parallel but it's mandatory to have a primary key on that table to support parallelism. If your table does not have a primary key then we can use –split-by argument to move our data from RDBMS to hdfs.

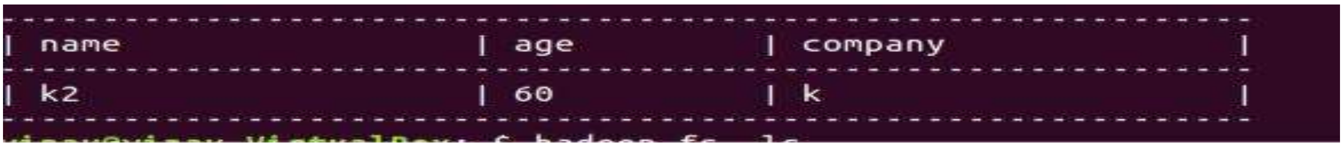


```
vinay@vinay-VirtualBox:~$ sqoop import-all-tables --connect jdbc:mysql://localho
st:3306/vinay --username root --password 123456 --m 1 --warehouse-dir /input/map
per
```

3) Eval is used to execute query using sqoop into database table and we can check at console.



```
vinay@vinay-VirtualBox:~$ sqoop eval --connect jdbc:mysql://localhost:3306/vinay
 --username root --password 123456 --query "select * from sqoop_demo1"
```



```
-----------------------------------------------------------------------
| name                      | age              | company                 |
-----------------------------------------------------------------------
| k2                        | 60               | k                       |
-----------------------------------------------------------------------
```

4) Export file kept in hdfs to mysql table

4.2 Sqoop - Transfer of traditional data of MySQL to HIVE

Hive is considered to be a data warehouse in the world of big data. It provides a platform to do MapReduce operations over RDMS. By Sqooping data can be imported from RDMS to Hive as well.

Import table from mysql to Hive .

sqoop import --connect jdbc:mysql://localhost:3306/reshu --username root --password 123456 --split-by Emp_id --columns Emp_id,name --table increment_load --target-dir increment_load6 --fields-terminated-by "," --hive-import --create-hive-table --hive-table default.increment_load --m 1

Note : mysql-connector-java.jar file should be present in lib folder of hive .

4.3 Sqoop - Transfer of traditional data of MySQL to HBase

HBase is a NoSQL Database i.e it is column distributed database. The model of HBase is similar to Goggle's Big Data. It has the capability of searching the data in a faster manner for a huge table.

4) Data import from mysql to Hbase :

In order to load data from mysql to Hbase either you have to create a table in hbase or you can add –hbase-create-table argument to create at run time.

# Create table in hbase : create 'Emp_info','Details'

sqoop command : sqoop import –connect jdbc:mysql://localhost/reshu –username root –password cloudera –table Emp_info –hbase-table Emp_info –column-family Details –habse-row-key Emp_id –m 1

Note: If table has more than1 column families and mysql table has multiple columns then same sqoop would be run with different column families. We can't mention more than column family in 1 sqoop command.

IV. Implementation Of Big data Tool Sqoop In Digital Library - Challanges

This section discuss about the challenges which digital libraries are facing while integrating the tool of latest technology like Sqoop for transfer of the data from one data base to another :-

1) Limited research of Big Data in the domain of Digital Library makes it difficult to make the Digital Library understand the relevance of the data transfer from traditional database management system (RDMS) to HDFS or HIVE or HBase.

2) In order to collect the data of individual user for customizing the Digital Library as per user's interest is not much appreciated yet.

3) Librarians are not much confident to migrate the system or the databases which they are using from so many years to a new platform with new functionality just because they are used to the traditional system.[5]

V. Conclusion And Future Action

6.1 Conclusion

In Digital Libraries, data is increasing day by day and now most of the researchers has declared the data of the digital library as Big Data. Digital Libraries are integrating themselves from repository of data to user driven database. Due to demand for the personalized demand of the users lots of data is generated like, users personal data, device data, cognitive style of the user to process the information., history of the data for getting the user's past interactions with the system, interest of the user, Domain expertise of the user, etc. All these data cannot be stored by traditional databases hence it has become really important for migrate the data of Digital Library from RDMS to Big Data. Sqoop import and export tool makes it very effective to transfer the data of one database to another database.

6.2 Future Action

The main focus of this research is to understand the importance of Big Data tool Sqoop for transferring the data of traditional database management system (RDBMS) to HDFS or Hive or HBase or on any other database system.

In future, we will work upon gathering the data by using following technology:-

1) Artificial Intelligence – It will help to search the relevant words used by the users for fetching the similar data over wide range of internet.

2) Spark Streaming – It will help to collect the real time data from the users and to convert them in top searches of a particular subject.

3) Big Data and Cloud – We can import and export the related data from the cloud quite easily.

Digital Libraries has huge scope in near future for its integration and modification.

6.3 Limitations

Implementation of new technologies like Big Data in Digital Library is still facing some issue due to following reasons :-

1) Working with Digital Library is not a point of interest for some researchers.
2) Librarians also have to upgrade themselves in order to use the latest technologies like Big Data.
3) Transferring such a huge data from traditional database to HDFS may face some implementations issues initially.
4) Transferring Legacy data in print format into digital format will include huge cost.

REFERENCES

[1] D. Vohra, Using apache sqoop, Pro Docker, Springer (2016), pp. 151-183
[2] A. Jain, Instant Apache Sqoop, Packt Publishing Ltd. (2013)
[3] https://www.guru99.com/introduction-to-flume-and-sqoop.html
[4] https://data-flair.training/blogs/sqoop-architecture-and-working/
[5] E.Frias-Martinez, G.Magoulas, S.Chen, R.Macredie. (2006). Automated user modeling for personalized digital libraries, Volume 26, Issue 3, June 2006, Pages 234-248