Big scholarly data

D.Kiranmayee[1], P Karthikeyan [2]

[1]Student, Dept. of information technology and engineering, Vellore institute of technology, Vellore, Tamilnadu, India.
[2]Professor, Dept. of information technology and engineering, Vellore institute of technology , Vellore ,Tamilnadu India.

---***---

Abstract: *Research progress and results are usually articulated through publishing articles. As a result of advancement in science, scientists around the world steadily produce a large volume of research articles, which provide the technological basis for worldwide dissemination of scientific findings [1]. In addition, researchers share their teaching materials such as slides and partial description of projects, patents and books through their homepages. The term Big Scholarly Data is coined for this rapidly growing scholarly source of information. Large collections of scholarly data have millions of authors, papers, citations, figures, tables, etc., as well as massive scale related data such as scholarly networks, digital libraries, etc. In this paper, present a survey of the emerging field of scholarly data. To the best of our knowledge, this paper is the first effort in providing a comprehensive review of scholarly data. Summary of the overall research issues on BSD from three Perspectives: scholarly data management, scholarly analysis methods, and scholarly Data applications. In the scholarly data management section, we review methods for scholarly data collections. Some popular digital libraries, academic search engines, and academic social networks are briefly introduced. At last discussion Regarding methods of investigating big scholarly data. This paper is subjected to provide a comprehensive understanding of research opportunities and challenges in the field of BSD and to find important issues for future explorative research*

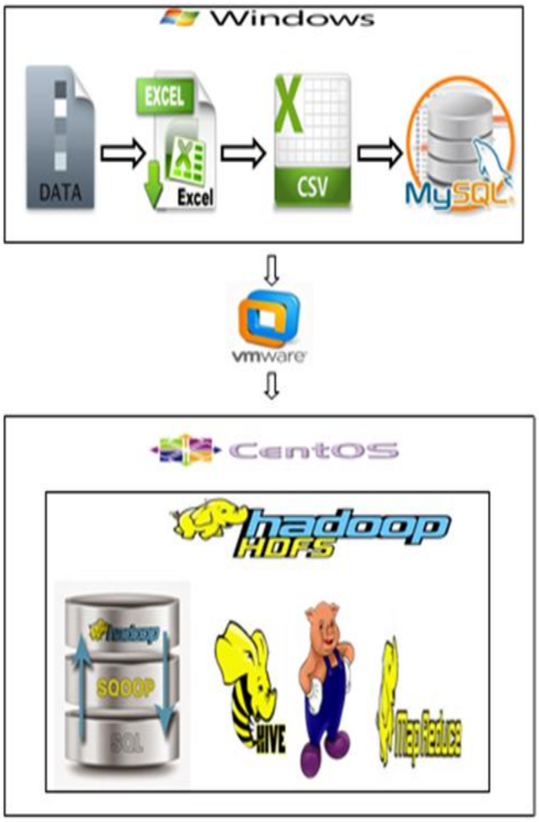
***Key Words*: BSD, Hadoop framework, Big data, sciences, Pig , Hive ,Sqoop**

1. INTRODUCTION

90% of the world's information was produced over the most recent couple of years [5]. Because of the appearance of new innovations, gadgets, and correspondence implies like informal communication destinations, the measure of information delivered by humankind is developing quickly consistently. The measure of information created by us from the earliest starting point of time till 2003 was 5 billion gigabytes. In the event that you heap up the information as plates it might fill a whole football field. A similar sum was made in each two days in 2011, and in like clockwork in 2013. This rate is as yet developing tremendously. Despite the fact that this data delivered is important and can be valuable when prepared, it is being dismissed. Enormous Data is an accumulation of huge datasets that can't be handled utilizing customary figuring techniques. In the field of huge information examination grouping of datasets turned into a testing issue [1]. It isn't a solitary strategy or an apparatus rather it includes numerous regions of business and innovation.

2. RELATED WORK:

This paper is about analyzing Scholarly data by using Hadoop tool along with some Hadoop ecosystems like hdfs, map reduce, sqoop, hive and pig. By using these tools processing of data without any limitation is possible and simple add number of machine to the cluster and we get results with very less time, no data lost problem, we can get high throughput, maintenance cost also very less and it is an open source software, it is compatible on all the platforms since it is Java based, joins, partitions and bucketing techniques are used in Hadoop. Hadoop framework which has overseen by an apache software foundation and it is used for storing and processing huge datasets with a cluster of commodity software. In scholarly data is related large volume of storage of research paper publishing website. when it comes to existing concept MySQL dB or rdbms it deals with providing backend by using MySQL which contains lots of drawbacks i.e. data limitation is just for about 6TB and processing time is high and if the data is lost we cannot recover and maintenance cost is very high and it is not an open source and there are lots of interoperability problems.so Hadoop tool is used to overcome this problem. VMware workstation is used for this project. VMware workstation is the best supporting software for big data and over this VMware workstation cent operating system can be installed this is the best supporting operating system for Hadoop framework, it improves performance of Hadoop .In the most first part the data will be included in the excel sheet in the form of fragments and all the data is converted into .csv file with a comma separated value. Excel sheet is supported by windows and it is not supported in different operating system i.e. mac operating system etc. so we use .csv file and then the data from .csv file is loaded in database table using sql. then by using sqoop migrating tool data is migrated from sql. to Hadoop.



All the data Is stored in hdfs and then data analysis is taken place by using pig and hive tools finally map reduce is used to avoid replications in the process. brief explanation is given below.

2.1Data storage in Existing Application (MySQL):

In MySQL is a relational database management system. RDBMS uses relations or tables to store Scholarly data as a matrix of rows by columns with primary keys and foreign keys. With MySQL language, Scholarly data in tables can be collected, stored, processed, retrieved, extracted and manipulated mostly for business purpose. Existing concept deals with providing backend by using MySQL which contains lot of drawbacks i.e. data limitation is that processing time is high when the data is huge and once data is lost we cannot recover so thus we proposing concept by using Hadoop tool.

Development TOOLS:

The java and Hadoop framework:

Huge Data is an accumulation of extensive datasets that can't be prepared utilizing conventional processing procedures. It isn't a solitary system or an apparatus; rather it includes numerous territories of business and innovation. Huge information includes the information created by various gadgets and applications. Given underneath are a portion of the fields that go under the umbrella of Big Data.

Discovery Data: It is a part of helicopter, planes, and streams, and so forth. It catches voices of the flight group, chronicles of mouthpieces and headphones, and the execution data of the flying machine.

Online networking Data: Social media, for example, Facebook and Twitter hold data and the perspectives posted by a huge number of individuals over the globe.

Stock Exchange Data: The stock trade information holds data about the 'purchase' and 'offer' choices made on an offer of various organizations made by the clients.

Power Grid Data: The power lattice information holds data devoured by a specific hub concerning a base station.

Transport Data: Transport information incorporates show, limit, separation and accessibility of a vehicle.

Web index Data: Search motors recover bunches of information from various databases.

Subsequently Big Data incorporates tremendous volume, high speed, and extensible assortment of information. The information in it will be of three kinds.

- Structured information: Relational information.
- Semi Structured information: XML information.
- Unstructured information: Word, PDF, Text, Media Logs.

Java is reasonable stage for huge information since; Java has been tried, refined, broadened, and demonstrated by a devoted group. What's more, numbering in excess of 6.5 million designers, it's the biggest and most dynamic on the planet. With its flexibility, effectiveness, and movability, Java has turned out to be important to designers by empowering them to:

- Write programming on one stage and run it on basically some other stage
- Create projects to keep running inside a Web program and Web administrations
- Develop server-side applications for online discussions, stores, surveys, HTML frames handling, and that's only the tip of the iceberg
- Combine applications or administrations utilizing the Java dialect to make exceedingly redid applications or administrations [3]
- Write effective and proficient applications for cell phones, remote processors, ease buyer items, and for all intents and purposes some other gadget with a computerized pulse

To be an Object Oriented dialect, any dialect must take after in any event the four qualities, for example, legacy, polymorphism, Encapsulation, dynamic authoritative.

The purpose for picking Hadoop system is Big Data is all over the place and there is right around a pressing need to gather and save whatever information is being created, for the dread of passing up a great opportunity for something vital. There is an immense measure of information skimming around. What we do with it is the only thing that is in any way important right at this point. This is the reason Big Data Analytics is in its outskirts. Enormous Data Analytics has turned out to be pivotal as it helps in enhancing business, choice makings and giving the greatest edge over the contenders. This applies for associations and also experts in the Analytics area. For experts, who are gifted in Big Data Analytics, there is a sea of chances out there.

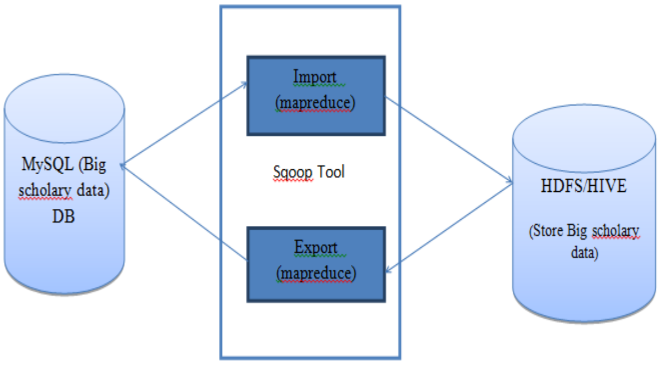
HDFS:

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Hadoop is colossally adaptable limit what's more, data dealing with system which supplements existing structures by dealing with data that is conventionally an issue for them [7]. It is a technique for securing enormous data sets transversely finished flowed gatherings of servers and a while later running "appropriated" dismemberment

Applications in every one gathering. It's expected to be vivacious, in that the Big Data applications will continue running really when frustrations occur in singular servers or gatherings [4]. Hadoop can in the meantime absorb and store any sort of data from a grouping of sources unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

2.2Connector (Sqoop):

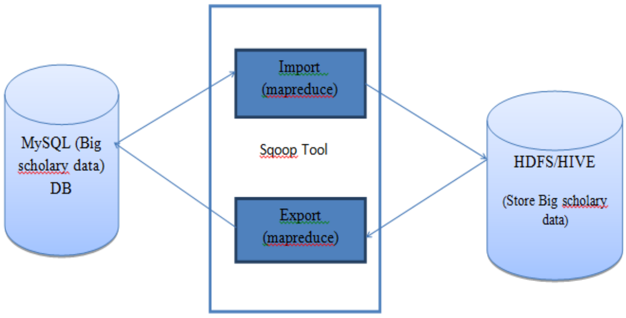
Sqoop is a command-line interface application for transferring Scholarly data between relational databases (MySQL) and Hadoop. Here in MySQL database having big scholarly data nothing but digital library data have to import it to HDFS using Swoop. Online Scholarly data can be moved into HDFS/Hive from MySQL and then it will generate the java classes. In previous cases, flow of data was from RDBMs to HDFS. Using "export" tool, we can import data from HDFS to RDBMs. Before performing export, Swoop fetches table metadata from MySQL database. Thus we first need to create a table with required metadata. The traditional application management system, that is, the interaction of applications with relational database using RDBMS, is one of the sources that generate Big Data. Such Big Data, generated by RDBMS, is stored in **Relational Database Servers** in the relational database structure



Data analysis:

2.3Analysis Query Language (Hive):

Hive is an information product house framework for Hadoop that runs SQL like questions called HQL (Hive inquiry dialect) which get inside changed over to delineate employments. In Hive, Scholarly information tables and databases are made first and afterward information is stacked into these tables. Hive as Scholarly information distribution Centre intended for overseeing and questioning just organized information that is put away in tables. Hive composes Scholarly information tables into allotments. It is a method for isolating a table into related parts in light of the estimations of parceled segments. Utilizing parcel, it is anything but difficult to question a bit of the given dataset. Tables or parcels are sub-separated into containers, to give additional structure to the Scholar data that might be utilized for more effective questioning. Bucketing works in light of the estimation of hash capacity of some section of a table. Hive: The term 'Enormous Data' is utilized for accumulations of vast datasets that incorporate colossal volume, high speed, and an assortment of information that is expanding step by step. Utilizing conventional information administration frameworks, it is hard to process Big Data. In this manner, the Apache Software Foundation acquainted a structure called Hadoop with illuminate Big Data administration and preparing challenges.



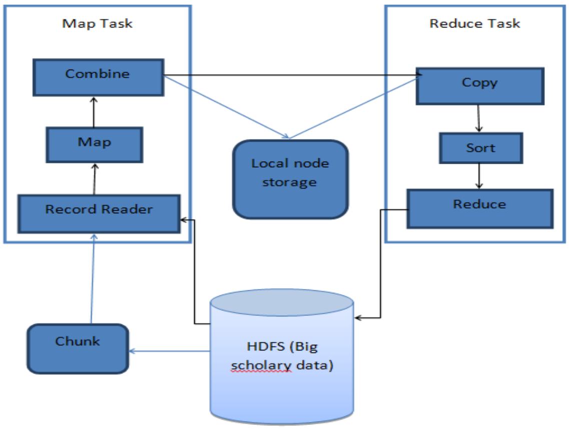
The Hadoop environment contains diverse sub-ventures (instruments, for example, Sqoop, Pig, and Hive that are utilized to help Hadoop modules. [4]

• Sqoop: It is utilized to import and fare information back and forth amongst HDFS and RDBMS.

• Hive: It is a stage used to create SQL writes contents to do Map Reduce tasks. Investigation Latin Script

2.4Data analysis (pig):

To examine Scholarly information utilizing Pig, software engineers need to compose contents utilizing Pig Latin dialect and execute them in intelligent mode utilizing the Grunt shell. Pig is an unusual state arrange for making Map Reduce ventures used with Hadoop. The tongue for this stage is called Pig Latin . Every one of these contents is inside changed over to Map and Reduce undertakings. In the wake of conjuring the Grunt shell, you can run your Pig contents in the shell. But LOAD and STORE, while playing out every other activity, Pig Latin proclamations take a connection as info and deliver another connection as yield. When you enter a Load explanation in the Grunt shell, its semantic looking freely is conveyed. To see the substance of the outline, you have to utilize the Dump administrator. Simply in the wake of playing out the landfill task, the MapReduce work for stacking the information into the document framework will be completed. Pig gives numerous inherent administrators to help information tasks like gathering, channels, requesting, and so forth. Apache Pig is an abnormal state information stream stage for execution Map Reduce projects of Hadoop. The dialect for Pig will be pig Latin.



The Pig contents get inside changed over to Map Reduce employments and get executed on information put away in HDFS. Each errand which can be accomplished utilizing PIG can likewise is accomplished utilizing java utilized as a part of Map decrease.

2.4MapReduce:

Map Reduce is a structure utilizing which we can compose applications to process tremendous measures of scholarly information, in parallel, on expansive groups of ware equipment in a dependable way. Map Reduce is a handling method and a program demonstrates for circulated registering in view of java. The Map Reduce calculation contains two critical undertakings, in particular Map and Reduce. Map Reduce program executes in three phases, to be specific guide organize, rearrange arrange, and lessen organize. The guide or mapper's activity is to process the information. By and large the information is as record or catalogue and is put away in the Hadoop document framework (HDFS). The information record is passed to the mapper work line by line. The mapper forms the information and makes a few little lumps of information. This stage is the blend of the Shuffle organizes and the Reduce arrange. The Reducer's activity is to process the information that originates from the mapper. Subsequent to preparing, it creates another arrangement of yield, which will be put away in the HDFS.

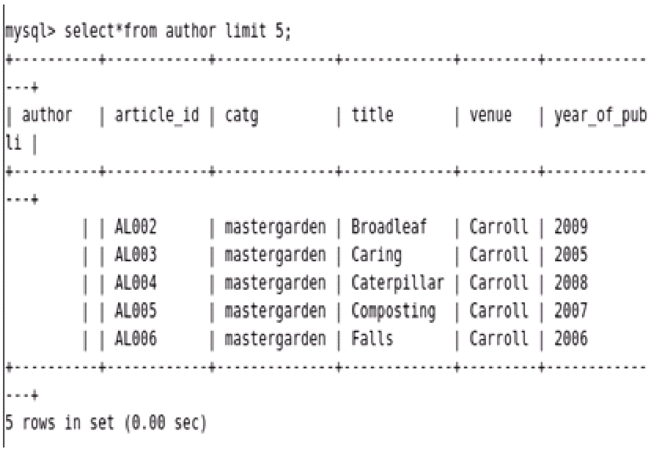
3. SCENARIO

3.1 Application

This paper is applied for student database and for searching author citations and so many other platforms .if we choose student database it is very easy to analyses the student performance and data is lost then it can retrieve the database too when it comes to hive and pig tools .hive is created by Facebook and hive query language is used in hive and it is query processing language used mostly for unstructured data .by using this hive tool it can analyze unstructured student data and it is used to query a portion of structured data and when it comes to pig it is created by yahoo and it is data flow language and it fits in pipeline paradigm and it is used for unstructured and structured and semi structured data.

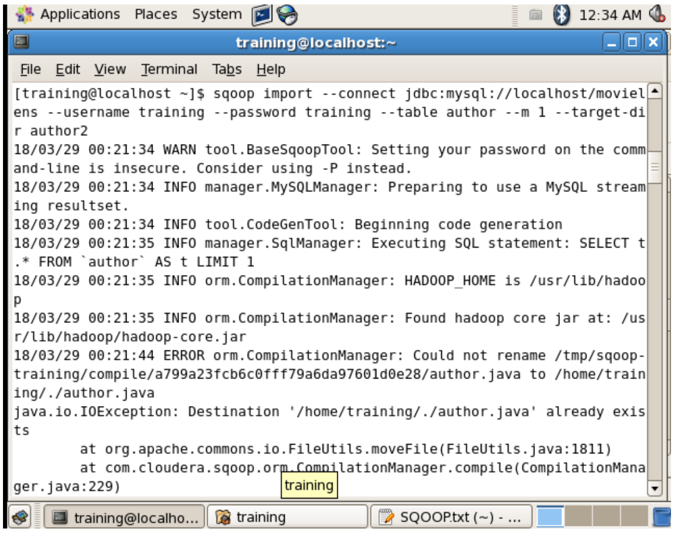
3.2 IMPLEMENTATION:

This is the author data retrieved through MySQL queries .



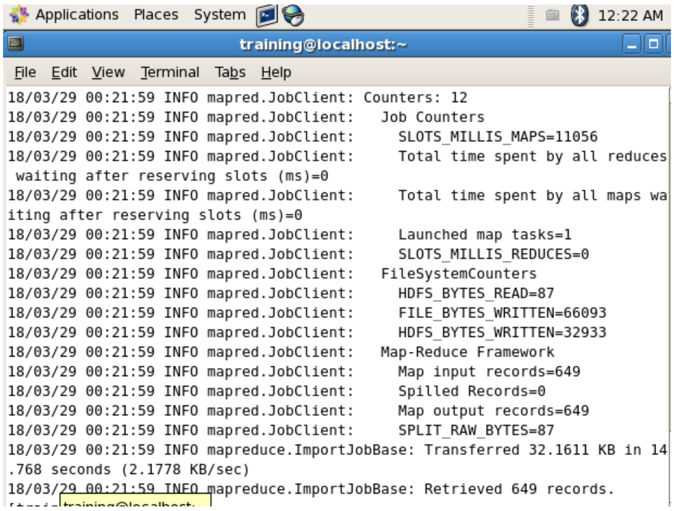
```
mysql> select*from author limit 5;
+----------+-----------+-------------+------------+---------+-----------
---+
| author   | article_id | catg         | title       | venue   | year_of_pub
li |
+----------+-----------+-------------+------------+---------+-----------
---+
         | | AL002      | mastergarden | Broadleaf   | Carroll | 2009
         | | AL003      | mastergarden | Caring      | Carroll | 2005
         | | AL004      | mastergarden | Caterpillar | Carroll | 2008
         | | AL005      | mastergarden | Composting  | Carroll | 2007
         | | AL006      | mastergarden | Falls       | Carroll | 2006
+----------+-----------+-------------+------------+---------+-----------
---+
5 rows in set (0.00 sec)
```

The below fig shows data migrated from MySQL to Hadoop using sqoop tool using sqoop queries.


Applications Places System
12:34 AM
training@localhost:~
File Edit View Terminal Tabs Help
training
training@localho... training SQOOP.txt (~) - ...

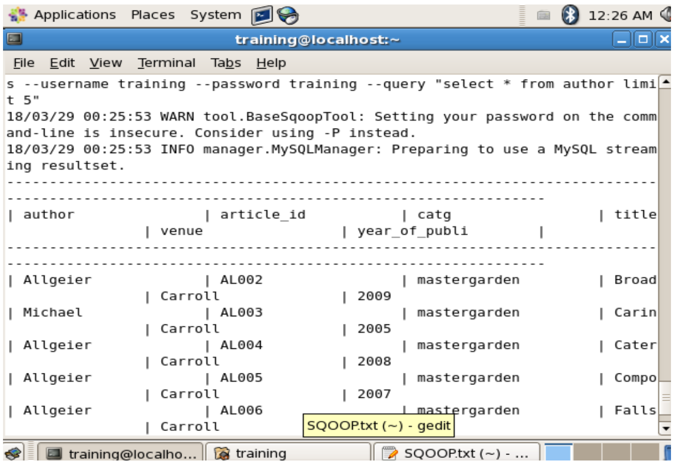
```
[training@localhost ~]$ sqoop import --connect jdbc:mysql://localhost/moviel
ens --username training --password training --table author --m 1 --target-di
r author2
18/03/29 00:21:34 WARN tool.BaseSqoopTool: Setting your password on the comm
and-line is insecure. Consider using -P instead.
18/03/29 00:21:34 INFO manager.MySQLManager: Preparing to use a MySQL stream
ing resultset.
18/03/29 00:21:34 INFO tool.CodeGenTool: Beginning code generation
18/03/29 00:21:35 INFO manager.SqlManager: Executing SQL statement: SELECT t
.* FROM `author` AS t LIMIT 1
18/03/29 00:21:35 INFO orm.CompilationManager: HADOOP_HOME is /usr/lib/hadoo
p
18/03/29 00:21:35 INFO orm.CompilationManager: Found hadoop core jar at: /us
r/lib/hadoop/hadoop-core.jar
18/03/29 00:21:44 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-
training/compile/a799a23fcb6c0fff79a6da97601d0e28/author.java to /home/train
ing/./author.java
java.io.IOException: Destination '/home/training/./author.java' already exis
ts
        at org.apache.commons.io.FileUtils.moveFile(FileUtils.java:1811)
        at com.cloudera.sqoop.orm.CompilationManager.compile(CompilationMana
ger.java:229)
```

The below fig shows the loaded table in hadoop using sqoop queries.



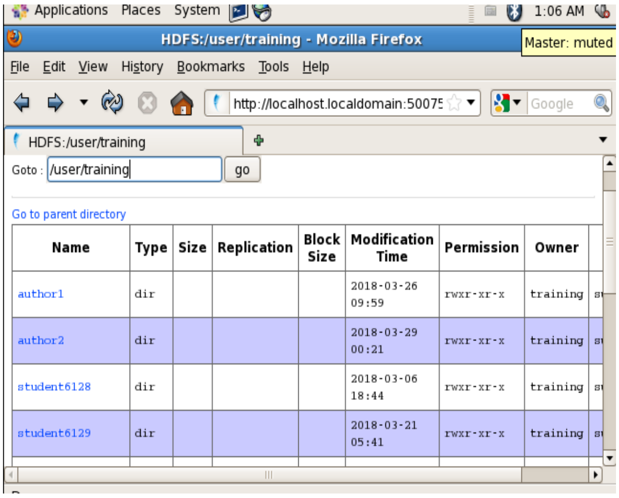
```
18/03/29 00:21:59 INFO mapred.JobClient: Counters: 12
18/03/29 00:21:59 INFO mapred.JobClient:   Job Counters
18/03/29 00:21:59 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=11056
18/03/29 00:21:59 INFO mapred.JobClient:     Total time spent by all reduces
 waiting after reserving slots (ms)=0
18/03/29 00:21:59 INFO mapred.JobClient:     Total time spent by all maps wa
iting after reserving slots (ms)=0
18/03/29 00:21:59 INFO mapred.JobClient:     Launched map tasks=1
18/03/29 00:21:59 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
18/03/29 00:21:59 INFO mapred.JobClient:   FileSystemCounters
18/03/29 00:21:59 INFO mapred.JobClient:     HDFS_BYTES_READ=87
18/03/29 00:21:59 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=66093
18/03/29 00:21:59 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=32933
18/03/29 00:21:59 INFO mapred.JobClient:   Map-Reduce Framework
18/03/29 00:21:59 INFO mapred.JobClient:     Map input records=649
18/03/29 00:21:59 INFO mapred.JobClient:     Spilled Records=0
18/03/29 00:21:59 INFO mapred.JobClient:     Map output records=649
18/03/29 00:21:59 INFO mapred.JobClient:     SPLIT_RAW_BYTES=87
18/03/29 00:21:59 INFO mapreduce.ImportJobBase: Transferred 32.1611 KB in 14
.768 seconds (2.1778 KB/sec)
18/03/29 00:21:59 INFO mapreduce.ImportJobBase: Retrieved 649 records.
```

The below fig shows the author dataset retrived through terminal using mysql queries.


Applications Places System
12:26 AM
training@localhost:~
File Edit View Terminal Tabs Help
SQOOP.txt (~) - gedit
training@localho... training SQOOP.txt (~) - ...

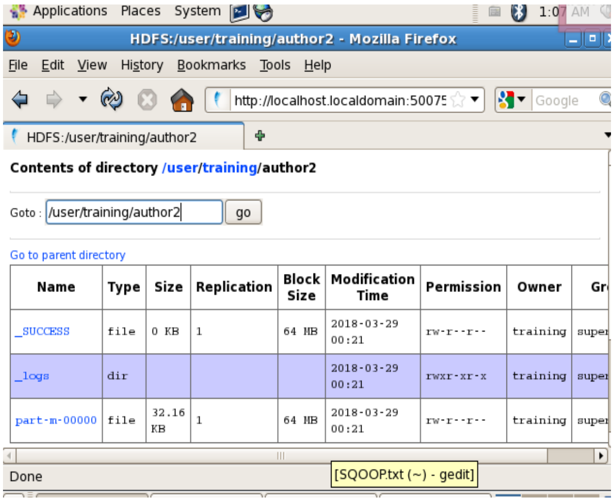
```
s --username training --password training --query "select * from author limi
t 5"
18/03/29 00:25:53 WARN tool.BaseSqoopTool: Setting your password on the comm
and-line is insecure. Consider using -P instead.
18/03/29 00:25:53 INFO manager.MySQLManager: Preparing to use a MySQL stream
ing resultset.
-----------------------------------------------------------------------------
-----------------------------------------------------------------
| author                | article_id             | catg                 | title
                | venue                  | year_of_publi        |
-----------------------------------------------------------------------------
-----------------------------------------------------------------
| Allgeier              | AL002                  | mastergarden         | Broad
                | Carroll                | 2009
| Michael               | AL003                  | mastergarden         | Carin
                | Carroll                | 2005
| Allgeier              | AL004                  | mastergarden         | Cater
                | Carroll                | 2008
| Allgeier              | AL005                  | mastergarden         | Compo
                | Carroll                | 2007
| Allgeier              | AL006                  | mastergarden         | Falls
                | Carroll
```

The below fig shows the author table 2 loaded in hdfs.


Applications Places System
1:06 AM
HDFS:/user/training - Mozilla Firefox
Master: muted
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training
Goto : /user/training go
Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	
author1	dir				2018-03-26 09:59	rwxr-xr-x	training	su
author2	dir				2018-03-29 00:21	rwxr-xr-x	training	su
student6128	dir				2018-03-06 18:44	rwxr-xr-x	training	su
student6129	dir				2018-03-21 05:41	rwxr-xr-x	training	su

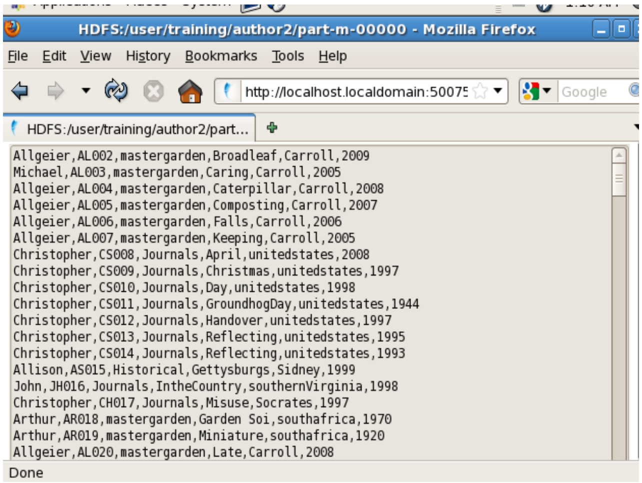
The below fig shows the content inside the directory which consists of success,logs and part number.


Applications Places System
1:07 AM
HDFS:/user/training/author2 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training/author2
Goto : /user/training/author2 go
Go to parent directory
Done
[SQOOP.txt (~) - gedit]

Contents of directory /user/training/author2

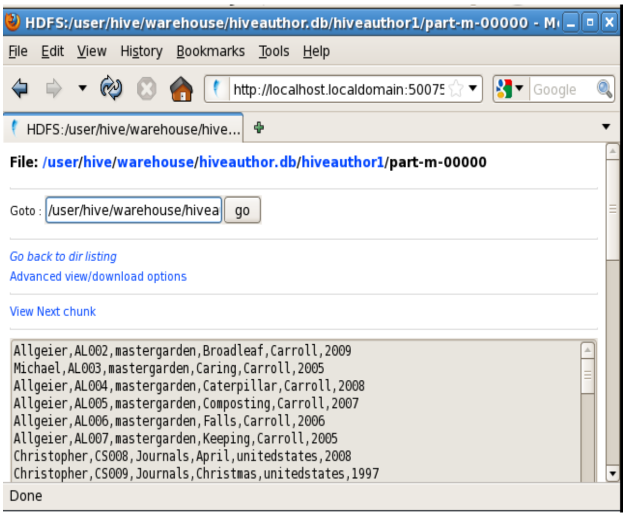
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Gr
_SUCCESS	file	0 KB	1	64 MB	2018-03-29 00:21	rw-r--r--	training	super
_logs	dir				2018-03-29 00:21	rwxr-xr-x	training	super
part-m-00000	file	32.16 KB	1	64 MB	2018-03-29 00:21	rw-r--r--	training	super

The below fig shows the result which consists inside the part number.


HDFS:/user/training/author2/part-m-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training/author2/part...
Done

```
Allgeier,AL002,mastergarden,Broadleaf,Carroll,2009
Michael,AL003,mastergarden,Caring,Carroll,2005
Allgeier,AL004,mastergarden,Caterpillar,Carroll,2008
Allgeier,AL005,mastergarden,Composting,Carroll,2007
Allgeier,AL006,mastergarden,Falls,Carroll,2006
Allgeier,AL007,mastergarden,Keeping,Carroll,2005
Christopher,CS008,Journals,April,unitedstates,2008
Christopher,CS009,Journals,Christmas,unitedstates,1997
Christopher,CS010,Journals,Day,unitedstates,1998
Christopher,CS011,Journals,GroundhogDay,unitedstates,1944
Christopher,CS012,Journals,Handover,unitedstates,1997
Christopher,CS013,Journals,Reflecting,unitedstates,1995
Christopher,CS014,Journals,Reflecting,unitedstates,1993
Allison,AS015,Historical,Gettysburgs,Sidney,1999
John,JH016,Journals,IntheCountry,southernVirginia,1998
Christopher,CH017,Journals,Misuse,Socrates,1997
Arthur,AR018,mastergarden,Garden Soi,southafrica,1970
Arthur,AR019,mastergarden,Miniature,southafrica,1920
Allgeier,AL020,mastergarden,Late,Carroll,2008
```

The below fig shows internal table should be created inside the dataware house which is presented in hive then data from hdfs loaded into the hive using part number.


HDFS:/user/hive/warehouse/hiveauthor.db/hiveauthor1/part-m-00000 - M
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/hive/warehouse/hive...
go
Done

File: /user/hive/warehouse/hiveauthor.db/hiveauthor1/part-m-00000

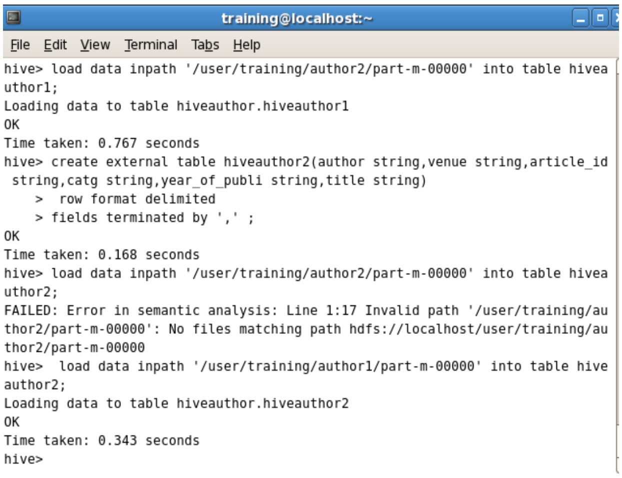
Goto : /user/hive/warehouse/hivea

Go back to dir listing
Advanced view/download options

View Next chunk

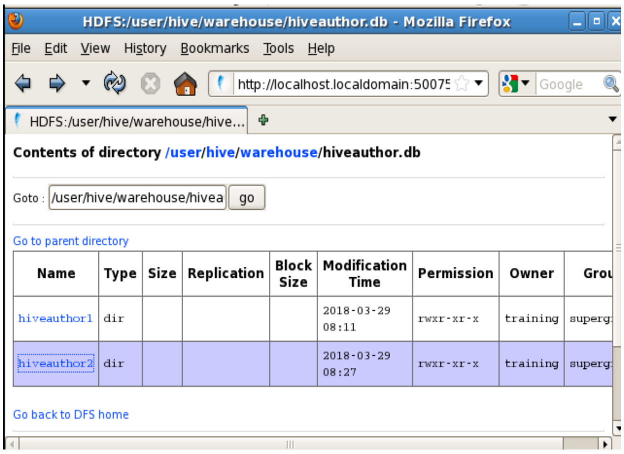
```
Allgeier,AL002,mastergarden,Broadleaf,Carroll,2009
Michael,AL003,mastergarden,Caring,Carroll,2005
Allgeier,AL004,mastergarden,Caterpillar,Carroll,2008
Allgeier,AL005,mastergarden,Composting,Carroll,2007
Allgeier,AL006,mastergarden,Falls,Carroll,2006
Allgeier,AL007,mastergarden,Keeping,Carroll,2005
Christopher,CS008,Journals,April,unitedstates,2008
Christopher,CS009,Journals,Christmas,unitedstates,1997
```

The below fig shows the external table can also be created in dataware house.



```
hive> load data inpath '/user/training/author2/part-m-00000' into table hivea
uthor1;
Loading data to table hiveauthor.hiveauthor1
OK
Time taken: 0.767 seconds
hive> create external table hiveauthor2(author string,venue string,article_id
 string,catg string,year_of_publi string,title string)
    >  row format delimited
    > fields terminated by ',' ;
OK
Time taken: 0.168 seconds
hive> load data inpath '/user/training/author2/part-m-00000' into table hivea
uthor2;
FAILED: Error in semantic analysis: Line 1:17 Invalid path '/user/training/au
thor2/part-m-00000': No files matching path hdfs://localhost/user/training/au
thor2/part-m-00000
hive>  load data inpath '/user/training/author1/part-m-00000' into table hive
author2;
Loading data to table hiveauthor.hiveauthor2
OK
Time taken: 0.343 seconds
hive>
```

The below fig shows external table.


HDFS:/user/hive/warehouse/hiveauthor.db - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/hive/warehouse/hive...
go

Contents of directory /user/hive/warehouse/hiveauthor.db

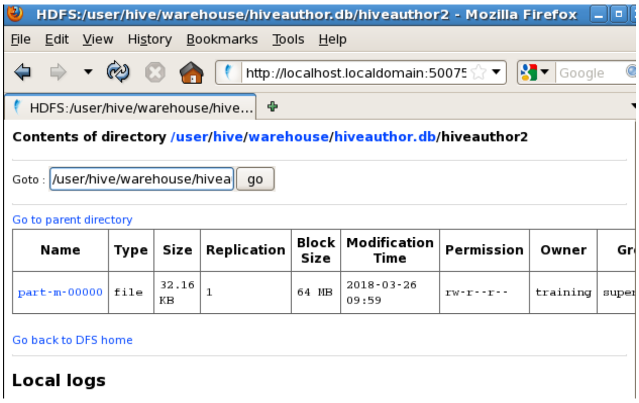
Goto : /user/hive/warehouse/hivea

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Grou
hiveauthor1	dir				2018-03-29 08:11	rwxr-xr-x	training	superg
hiveauthor2	dir				2018-03-29 08:27	rwxr-xr-x	training	superg

Go back to DFS home

The below fig shows the data loaded in external number by partnumber.


HDFS:/user/hive/warehouse/hiveauthor.db/hiveauthor2 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/hive/warehouse/hive...
go

Contents of directory /user/hive/warehouse/hiveauthor.db/hiveauthor2

Goto : /user/hive/warehouse/hivea

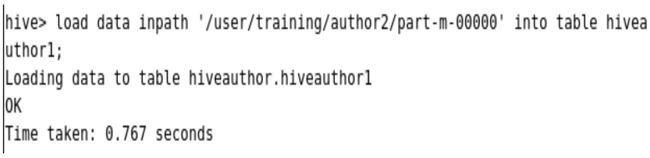
Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Gr
part-m-00000	file	32.16 KB	1	64 MB	2018-03-26 09:59	rw-r--r--	training	super

Go back to DFS home

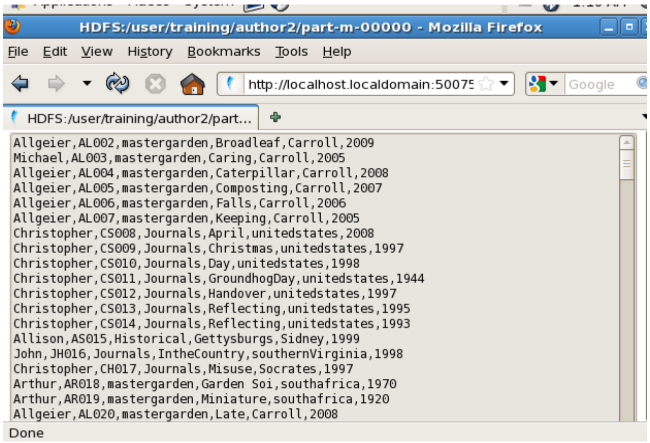
Local logs

The below query used to load the dataset into hive



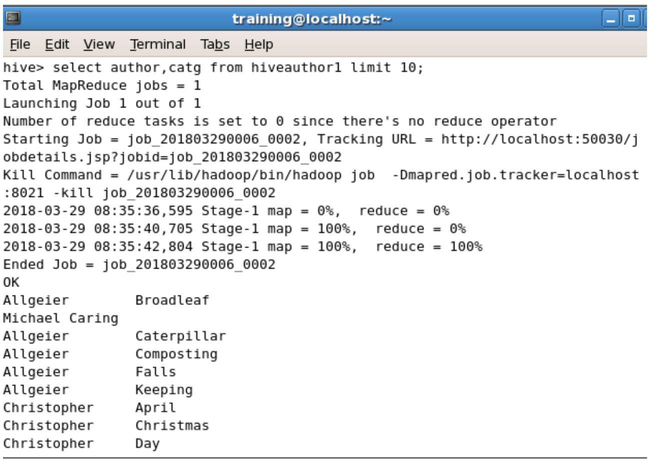
```
hive> load data inpath '/user/training/author2/part-m-00000' into table hivea
uthor1;
Loading data to table hiveauthor.hiveauthor1
OK
Time taken: 0.767 seconds
```

The below fig shows the final result present inside the partnumber.


HDFS:/user/training/author2/part-m-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training/author2/part...
Done

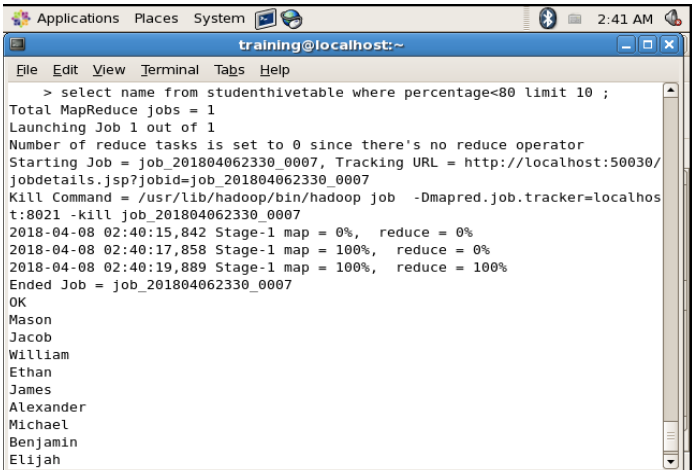
```
Allgeier,AL002,mastergarden,Broadleaf,Carroll,2009
Michael,AL003,mastergarden,Caring,Carroll,2005
Allgeier,AL004,mastergarden,Caterpillar,Carroll,2008
Allgeier,AL005,mastergarden,Composting,Carroll,2007
Allgeier,AL006,mastergarden,Falls,Carroll,2006
Allgeier,AL007,mastergarden,Keeping,Carroll,2005
Christopher,CS008,Journals,April,unitedstates,2008
Christopher,CS009,Journals,Christmas,unitedstates,1997
Christopher,CS010,Journals,Day,unitedstates,1998
Christopher,CS011,Journals,GroundhogDay,unitedstates,1944
Christopher,CS012,Journals,Handover,unitedstates,1997
Christopher,CS013,Journals,Reflecting,unitedstates,1995
Christopher,CS014,Journals,Reflecting,unitedstates,1993
Allison,AS015,Historical,Gettysburgs,Sidney,1999
John,JH016,Journals,IntheCountry,southernVirginia,1998
Christopher,CH017,Journals,Misuse,Socrates,1997
Arthur,AR018,mastergarden,Garden Soi,southafrica,1970
Arthur,AR019,mastergarden,Miniature,southafrica,1920
Allgeier,AL020,mastergarden,Late,Carroll,2008
```

The below fig shows the analysis of dataset using hive.



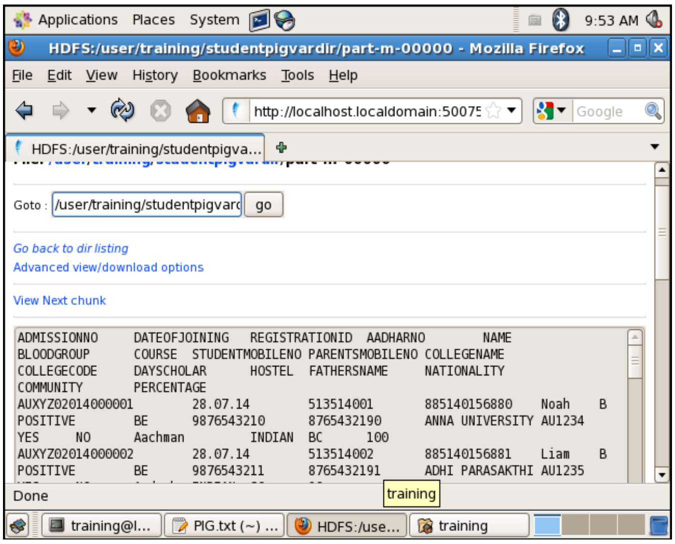
```
hive> select author,catg from hiveauthor1 limit 10;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201803290006_0002, Tracking URL = http://localhost:50030/j
obdetails.jsp?jobid=job_201803290006_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost
:8021 -kill job_201803290006_0002
2018-03-29 08:35:36,595 Stage-1 map = 0%,  reduce = 0%
2018-03-29 08:35:40,705 Stage-1 map = 100%,  reduce = 0%
2018-03-29 08:35:42,804 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201803290006_0002
OK
Allgeier        Broadleaf
Michael Caring
Allgeier        Caterpillar
Allgeier        Composting
Allgeier        Falls
Allgeier        Keeping
Christopher     April
Christopher     Christmas
Christopher     Day
```

The below fig shows the analysis of another dataset using hive.



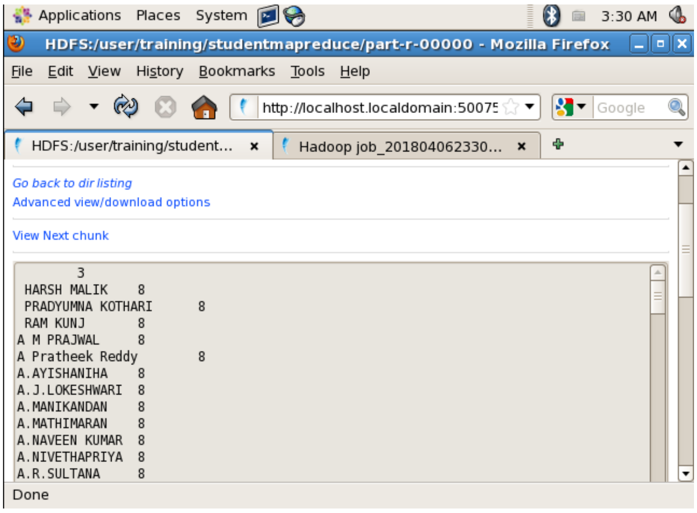
```
    > select name from studenthivetable where percentage<80 limit 10 ;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201804062330_0007, Tracking URL = http://localhost:50030/
jobdetails.jsp?jobid=job_201804062330_0007
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhos
t:8021 -kill job_201804062330_0007
2018-04-08 02:40:15,842 Stage-1 map = 0%,  reduce = 0%
2018-04-08 02:40:17,858 Stage-1 map = 100%,  reduce = 0%
2018-04-08 02:40:19,889 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201804062330_0007
OK
Mason
Jacob
William
Ethan
James
Alexander
Michael
Benjamin
Elijah
```

The below fig shows the analysis of dataset using pig.


Applications Places System
9:53 AM
HDFS:/user/training/studentpigvardir/part-m-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training/studentpigva...
Goto : /user/training/studentpigvard
go
Go back to dir listing
Advanced view/download options
View Next chunk
Done
training
training@l...
PIG.txt (~) ...
HDFS:/use...
training

```
ADMISSIONNO     DATEOFJOINING   REGISTRATIONID  AADHARNO        NAME
BLOODGROUP      COURSE  STUDENTMOBILENO PARENTSMOBILENO COLLEGENAME
COLLEGECODE     DAYSCHOLAR      HOSTEL  FATHERSNAME     NATIONALITY
COMMUNITY       PERCENTAGE
AUXYZ02014000001        28.07.14        513514001       885140156880    Noah    B
POSITIVE        BE      9876543210      8765432190      ANNA UNIVERSITY AU1234
YES     NO      Aachman         INDIAN  BC      100
AUXYZ02014000002        28.07.14        513514002       885140156881    Liam    B
POSITIVE        BE      9876543211      8765432191      ADHI PARASAKTHI AU1235
```

The below fig shows the analysis of dataset using mapreduce to avoid replications.


Applications Places System
3:30 AM
HDFS:/user/training/studentmapreduce/part-r-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075
Google
HDFS:/user/training/student...
Hadoop job_201804062330...
Go back to dir listing
Advanced view/download options
View Next chunk
Done

```
        3
 HARSH MALIK    8
 PRADYUMNA KOTHARI      8
 RAM KUNJ       8
A M PRAJWAL     8
A Pratheek Reddy        8
A.AYISHANIHA    8
A.J.LOKESHWARI  8
A.MANIKANDAN    8
A.MATHIMARAN    8
A.NAVEEN KUMAR  8
A.NIVETHAPRIYA  8
A.R.SULTANA     8
```

pseudo code:

```
[ import packages conf. configured, fs.Path, io.IntWritable, io.Text, mapreduce.Job,

mapreduce.lib.input.FileInputFormat, mapreduce.lib.output.FileOutputFormat

 create a public class Pop_Driver extending Configured
 create a main string function with Exception
  if string length is not equal to 2
  print “please give proper i/p and o/p dir”
 create a job
set pop_Driver.class within JarByClass
FileInputFormat.setInputPaths (new Path (args [0]));
 FileOutputFormat.setOutputPath (j, new Path (args [1]));
Set mapper class
Set reducer class
Set map output key class for text class
Set map output value class for IntWritable
Set output key class for text class
Set output value class for IntWritable]
```

The above pseudo code is run in eclipse. it is the best software used for map reduce. Above queries can be run in terminal and final output is viewed in browser. We used 50070 configuration to connect the browser .The data stored in hdfs can be viewed through browser.

4. CONCLUSION:

This paper, presented a study on scholarly data and prediction regarding research paper about websites, author, co- author, citation. To analysis the scholarly data in Hadoop ecosystem [2]. Hadoop ecosystem is having hive, pig, map reduce tools for processing whether output will take less time to process and result will be very fast. Hence in this project already scholarly data which is traditionally going to store in RDBMS going to less performance hence by using Hadoop tool faster and efficiently processing the data.

5. FUTURE ENHANCEMENT:

Apache Spark is an open source preparing motor worked around speed, instance of utilization, and investigation. In the event that you have a lot of information that requires low idleness handling that a run of the mill Map Reduce program can't give, Spark is the option. Start gives in-memory group processing to exceptionally quick speed and backings Java, Scales, and Python APIs for simplicity of improvement. [6]

6. REFERENCES:

[1]C. Carnage, J. Wu, K. Williams, S. Das, M. Khabsa, P. Teregowda, and C. L. Giles, "Automatic identification of research articles from crawled documents," Proceedings of WSDM-WSCBD, 2014.

[2]Sara land set taghi M.khoshgoftaar, Aaron Tawfiq Hasanin "a survey of open source tools for machine learning in big data with Hadoop ecosystem" 2015 [3]AntonioCarzaniga, Alessandra Gorla, Andrea Mattavelli, Nicolo Perino "self-healing technique for java applications"2012

[4] Manoj Kumar Singh, Dr.Parveen Kumar University, Gajraula, U.P, India "Hadoop: A Big Data Management Framework for Storage, Scalability, Complexity, Distributed Files and Processing of Massive Datasets".2014

[5] "Big Data, for better or worse: 90% of world's data generated over last two years" source: SINTEF 2013.

[6]"apache spark vs. map reduces" data flair 2016

[7] Intel It Centre, "Planning Guide- Getting started with Big Data"