An Analysis of Summer Olympics Using Open Source Big Data Technologies

Shikhar Srivastava
x18106960

***Abstract*—Game events which occur on a global scale are important events in which players from all over the globe participate. Players participate in many different sports which are scheduled to happen over the course of many days. Olympics is one of the major sport events which occur after every four years in which many athletes participate from all over the world. This mega game event is further bifurcated in two Olympic events i.e. Summer Olympics and Winter Olympics with just a single difference that all the winter sports like skiing are scheduled to happen in Winter Olympics. It would be very interesting to analyze data recorded for this majestic game event and look at some insights. This project uses open source big data technologies like MapReduce, Sqoop, HBase, Hive, Pig and HDFS and tries to extract few useful and insightful information. This project also follows an approach that how these open source big data technologies can be efficiently used to store and process data which in our case is Olympic dataset.**

***Index Terms*—MapReduce, MySQL, Sqoop, HBase, Hive, Pig, Hadoop, HDFS, Olympics data, R, Tableau**

I. Introduction

The world is witnessing an era of data explosion, an era where it is a known fact that data is a very important asset. As the volume of data is increasing exponentially, it became a need and necessity to be able to store and process this huge volume of data efficiently. With this explosion in data, new approaches emerged to store it, where people started to store huge volume data on distributed platforms rather than storing it in traditional relational databases. Many technologies also started to emerge which could manage data on a distributed platform. This project utilizes some of these popularly known open source technologies to analyze Olympic dataset and aims to showcase how these technologies can be leveraged to store and process huge datasets.

International sport events occur at a very large scale involving a large number of participants from many different countries. These participants compete in several events across many sports. Olympic games are one of the most popularly known sport events which takes place after every four years. Summer and Winter Olympics are the two types of Olympic events which are held where Winter Olympics are especially organized for winter sports. This project utilizes Olympic dataset which is downloaded from Kaggle [1] which contains player wise details like in which sport the player participated, in which Olympic, in which sport, nationality of the player, gender and few more details. Columns and their brief description is given in the appendix section of this document. The Olympic dataset is fed in different databases like MySQL, HBase, Hive, Pig and is then further analyzed by building queries. The objective of this project is to feed Olympic dataset to above mentioned databases and explore data to find some key insights, which are explained in the methodology section.

The rest of the document is structured in four different sections: Related Work, a concise literature review on relevant research papers; Methodology, explains all the steps followed for implementation of this project; Results, shows and explains results from the implementation.

II. Related Work

Olympics are ancient sport events which roughly started in around 5th century B.C. [2] when intellectuals known as Sophists travelled to Olympia to demonstrate combative oratory. This reflects that Olympics have rich history which dates back many hundreds of years. Major role in reviving the ancient Olympic event was played by Pierre de Coubertin [3] who provided a launch pad for modern Olympics. As the roots of Olympics come Greece, it faced negligence when the 1896 Olympic games were about to start. His name was no where mentioned and the greeks accused him of being a thief to steal their inheritance. Olympics are not just bound to happen in Greece, but now many other countries host Olympic events every four years. It greatly impacts the economic situation of host countries. A research has also been done to examine and evaluate methods and assumptions used by economic studies on Olympic host countries. Olympic games not just impact economy of host countries but they also have a social impact on which a research has been done for based on Sydney Olympics [4].

This project tries to analyze Olympic data to look at some insightful information. There are many open source big data technologies available to store, manage and process large scale data like MySQL, HBase, Hive, Pig etc. There have been many researches on these distributed platforms. One such research evaluated the performance of HBase on random reads and writes and further does a comparative study of HBase on HDFS and MySQL on HDFS. Similarly, Hive is also a famous platform for big data processing on top of HDFS. [5] provides a descriptive explanation of Hive architecture and why is it suitable to be used for large scale data on HDFS. Hive also comes with almost SQL like querying language which is called HiveQL. Pig is another popular big data technology which is widely used to store and manage high volume of data. A new querying language was designed to operate on Pig by a group of researchers [6] and was called as Pig Latin which has the declarative style of SQL and procedural style of MapReduce. Storing, managing and processing of large data can be efficiently handled by above mentioned big data technologies, but what if there is a requirement to move data from RDBMS to hadoop ecosystem. Sqoop is very a robust framework which is widely used to transfer data from relational database like MySQL to HDFS. Sqoop works on JDBC drivers to transfer data into hadoop [7] [8]

III. Methodology

This section provides a detailed explanation of data used in this project along with steps followed to prepare, store and analyze Olympic data.

A. Data Acquisition and Preprocessing

Dataset used in this project has been downloaded from Kaggle website [1] which is a public dataset, so there are no ethical issues associated with dataset used. Although on above mentioned link there are two datasets in the form of CSV files which includes athelete_events.csv and noc_regions.csv, but for this project we are only using athelete_events.csv. CSV file of size 39.58 MB containing all player wise Olympic data was downloaded.

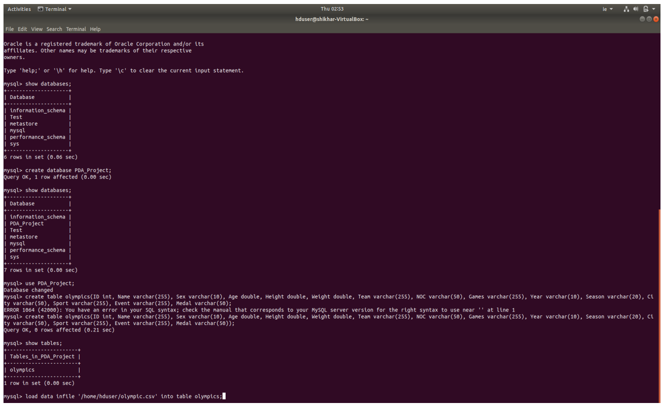
The downloaded dataset was in a pretty good condition and didn't need any cleaning as such. The only thing which was carried out as part of pre-processing was to replace NULL values in Medal column with 'No Medal' and remove all NAs from the dataset. This two step re-processing task was done in R by making use of tidyr library of R, (2_Step_Pre_Processing.R) is present in the compressed folder attached. The output file from this code was then copied and pasted in shared folder to use it in Ubuntu VM.

B. Step by Step Process Followed

This section details all the steps followed to store and process Olympic data in different databases.

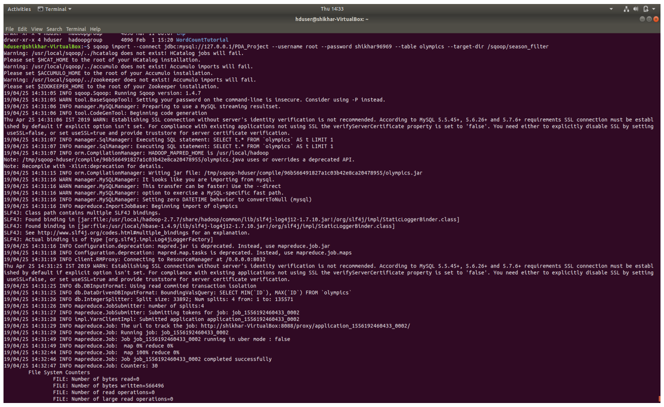
Step 1: File copied to shared folder was first again copied to hduser just to make sure that there are no conflicts with permissions. After this, MySQL was launched in which a new database was created named 'PDA_Project' in which a new table 'olympics' was created and data from the Olympic csv file was then loaded in this table. Screenshot in Fig. 1.

Step 2: 'olympics' table data stored in MySQL was then stored on HDFS to execute MapReduce shell script on this dataset. MySQL table data was copied to HDFS with the help of Sqoop which stored MySQL table data in sqoop/season_filter directory in HDFS. Fig. 2 and Fig. 3 shows the successful execution of sqoop command.


Activities
Terminal
Thu 02:53
ie
hduser@shikhar-VirtualBox: ~
File Edit View Search Terminal Help

```
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| Test               |
| metastore          |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
6 rows in set (0.06 sec)

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

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| PDA_Project        |
| Test               |
| metastore          |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
7 rows in set (0.00 sec)

mysql> use PDA_Project;
Database changed
mysql> create table olympics(ID int, Name varchar(255), Sex varchar(10), Age double, Height double, Weight double, Team varchar(255), NOC varchar(50), Games varchar(255), Year varchar(10), Season varchar(20), Ci
ty varchar(50), Sport varchar(255), Event varchar(255), Medal varchar(50);
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '' at line 1
mysql> create table olympics(ID int, Name varchar(255), Sex varchar(10), Age double, Height double, Weight double, Team varchar(255), NOC varchar(50), Games varchar(255), Year varchar(10), Season varchar(20), Ci
ty varchar(50), Sport varchar(255), Event varchar(255), Medal varchar(50));
Query OK, 0 rows affected (0.21 sec)

mysql> show tables;
+-----------------------+
| Tables_in_PDA_Project |
+-----------------------+
| olympics              |
+-----------------------+
1 row in set (0.00 sec)

mysql> load data infile '/home/hduser/olympic.csv' into table olympics;
```

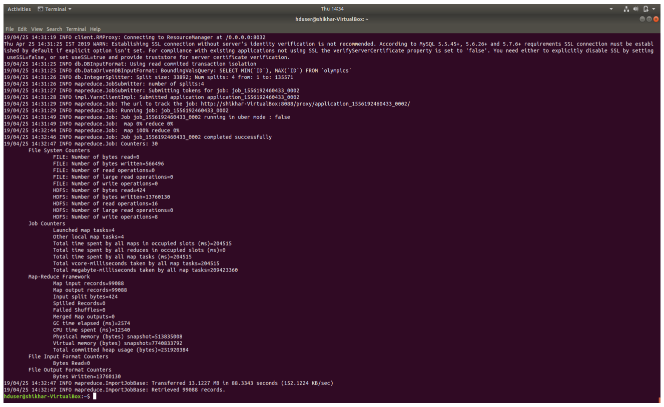
Fig. 1. Loading Data in MySQL


Activities
Terminal
Thu 14:33
hduser@shikhar-VirtualBox: ~
File Edit View Search Terminal Help

```
drwxr-xr-x 4 hduser  hadoopgroup      4096 Feb  1 15:20 WordCountTutorial
hduser@shikhar-VirtualBox:~$ sqoop import --connect jdbc:mysql://127.0.0.1/PDA_Project --username root --password shikhar96969 --table olympics --target-dir /sqoop/season_filter
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
19/04/25 14:31:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
19/04/25 14:31:05 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
19/04/25 14:31:06 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/04/25 14:31:06 INFO tool.CodeGenTool: Beginning code generation
Thu Apr 25 14:31:06 IST 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be establ
ished by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting
 useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
19/04/25 14:31:07 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `olympics` AS t LIMIT 1
19/04/25 14:31:07 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `olympics` AS t LIMIT 1
19/04/25 14:31:07 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/96b566491827a1c03b42e8ca20478955/olympics.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/04/25 14:31:15 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/96b566491827a1c03b42e8ca20478955/olympics.jar
19/04/25 14:31:16 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/04/25 14:31:16 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/04/25 14:31:16 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/04/25 14:31:16 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/04/25 14:31:16 INFO mapreduce.ImportJobBase: Beginning import of olympics
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
19/04/25 14:31:16 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/04/25 14:31:18 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/04/25 14:31:19 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
Thu Apr 25 14:31:25 IST 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be establ
ished by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting
 useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
19/04/25 14:31:25 INFO db.DBInputFormat: Using read commited transaction isolation
19/04/25 14:31:25 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`ID`), MAX(`ID`) FROM `olympics`
19/04/25 14:31:26 INFO db.IntegerSplitter: Split size: 33892; Num splits: 4 from: 1 to: 135571
19/04/25 14:31:26 INFO mapreduce.JobSubmitter: number of splits:4
19/04/25 14:31:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1556192460433_0002
19/04/25 14:31:28 INFO impl.YarnClientImpl: Submitted application application_1556192460433_0002
19/04/25 14:31:29 INFO mapreduce.Job: The url to track the job: http://shikhar-VirtualBox:8088/proxy/application_1556192460433_0002/
19/04/25 14:31:29 INFO mapreduce.Job: Running job: job_1556192460433_0002
19/04/25 14:31:49 INFO mapreduce.Job: Job job_1556192460433_0002 running in uber mode : false
19/04/25 14:31:49 INFO mapreduce.Job:  map 0% reduce 0%
19/04/25 14:32:44 INFO mapreduce.Job:  map 100% reduce 0%
19/04/25 14:32:46 INFO mapreduce.Job: Job job_1556192460433_0002 completed successfully
19/04/25 14:32:47 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566496
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

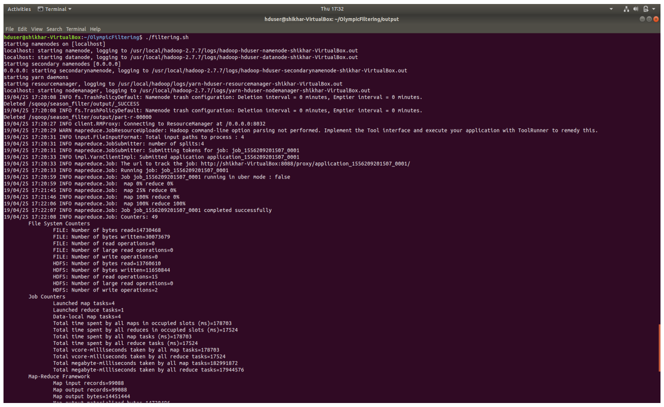
Fig. 2. Sqoop Output a

Step 3: This step involves filtering Olympic data for just summer Olympics in Season column. For this a MapReduce java program was written in eclipse and jar file was exported to


Activities
Terminal
Thu 14:34
hduser@shikhar-VirtualBox: ~
File Edit View Search Terminal Help

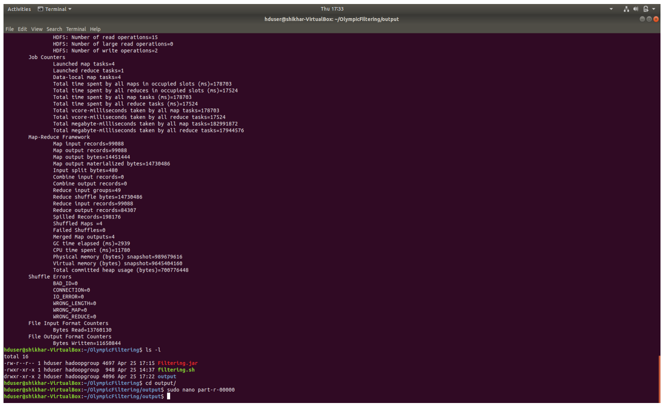
```
19/04/25 14:31:19 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
Thu Apr 25 14:31:25 IST 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be establ
ished by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting
 useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
19/04/25 14:31:25 INFO db.DBInputFormat: Using read commited transaction isolation
19/04/25 14:31:25 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`ID`), MAX(`ID`) FROM `olympics`
19/04/25 14:31:26 INFO db.IntegerSplitter: Split size: 33892; Num splits: 4 from: 1 to: 135571
19/04/25 14:31:26 INFO mapreduce.JobSubmitter: number of splits:4
19/04/25 14:31:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1556192460433_0002
19/04/25 14:31:28 INFO impl.YarnClientImpl: Submitted application application_1556192460433_0002
19/04/25 14:31:29 INFO mapreduce.Job: The url to track the job: http://shikhar-VirtualBox:8088/proxy/application_1556192460433_0002/
19/04/25 14:31:29 INFO mapreduce.Job: Running job: job_1556192460433_0002
19/04/25 14:31:49 INFO mapreduce.Job: Job job_1556192460433_0002 running in uber mode : false
19/04/25 14:31:49 INFO mapreduce.Job:  map 0% reduce 0%
19/04/25 14:32:44 INFO mapreduce.Job:  map 100% reduce 0%
19/04/25 14:32:46 INFO mapreduce.Job: Job job_1556192460433_0002 completed successfully
19/04/25 14:32:47 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=566496
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=424
                HDFS: Number of bytes written=13760130
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=204515
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=204515
                Total vcore-milliseconds taken by all map tasks=204515
                Total megabyte-milliseconds taken by all map tasks=209423360
        Map-Reduce Framework
                Map input records=99088
                Map output records=99088
                Input split bytes=424
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2574
                CPU time spent (ms)=12540
                Physical memory (bytes) snapshot=513835008
                Virtual memory (bytes) snapshot=7740833792
                Total committed heap usage (bytes)=251920384
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=13760130
19/04/25 14:32:47 INFO mapreduce.ImportJobBase: Transferred 13.1227 MB in 88.3343 seconds (152.1224 KB/sec)
19/04/25 14:32:47 INFO mapreduce.ImportJobBase: Retrieved 99088 records.
hduser@shikhar-VirtualBox:~$
```

Fig. 3. Sqoop Output b


Activities
Terminal
Thu 17:32
hduser@shikhar-VirtualBox: ~/OlympicFiltering/output
File Edit View Search Terminal Help

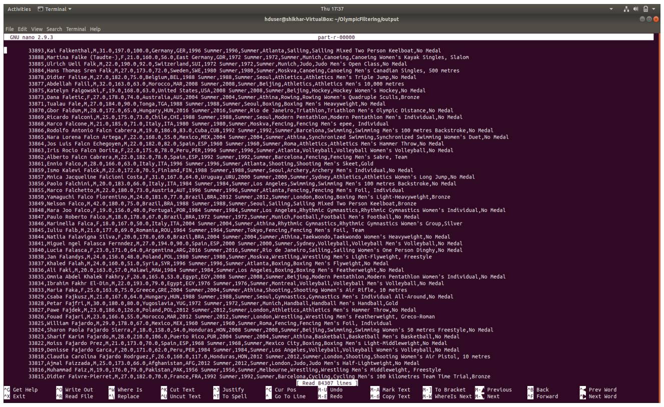
```
hduser@shikhar-VirtualBox:~/OlympicFiltering$ ./filtering.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.7.7/logs/hadoop-hduser-namenode-shikhar-VirtualBox.out
localhost: starting datanode, logging to /usr/local/hadoop-2.7.7/logs/hadoop-hduser-datanode-shikhar-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.7.7/logs/hadoop-hduser-secondarynamenode-shikhar-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-shikhar-VirtualBox.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.7.7/logs/yarn-hduser-nodemanager-shikhar-VirtualBox.out
19/04/25 17:20:08 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /sqoop/season_filter/output/_SUCCESS
19/04/25 17:20:08 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /sqoop/season_filter/output/part-r-00000
19/04/25 17:20:27 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/25 17:20:29 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/04/25 17:20:31 INFO input.FileInputFormat: Total input paths to process : 4
19/04/25 17:20:31 INFO mapreduce.JobSubmitter: number of splits:4
19/04/25 17:20:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1556209201507_0001
19/04/25 17:20:33 INFO impl.YarnClientImpl: Submitted application application_1556209201507_0001
19/04/25 17:20:33 INFO mapreduce.Job: The url to track the job: http://shikhar-VirtualBox:8088/proxy/application_1556209201507_0001/
19/04/25 17:20:33 INFO mapreduce.Job: Running job: job_1556209201507_0001
19/04/25 17:20:59 INFO mapreduce.Job: Job job_1556209201507_0001 running in uber mode : false
19/04/25 17:20:59 INFO mapreduce.Job:  map 0% reduce 0%
19/04/25 17:21:45 INFO mapreduce.Job:  map 25% reduce 0%
19/04/25 17:21:46 INFO mapreduce.Job:  map 100% reduce 0%
19/04/25 17:22:06 INFO mapreduce.Job:  map 100% reduce 100%
19/04/25 17:22:07 INFO mapreduce.Job: Job job_1556209201507_0001 completed successfully
19/04/25 17:22:08 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=14730468
                FILE: Number of bytes written=30073679
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=13760610
                HDFS: Number of bytes written=11650844
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=178703
                Total time spent by all reduces in occupied slots (ms)=17524
                Total time spent by all map tasks (ms)=178703
                Total time spent by all reduce tasks (ms)=17524
                Total vcore-milliseconds taken by all map tasks=178703
                Total vcore-milliseconds taken by all reduce tasks=17524
                Total megabyte-milliseconds taken by all map tasks=182991872
                Total megabyte-milliseconds taken by all reduce tasks=17944576
        Map-Reduce Framework
                Map input records=99088
                Map output records=99088
                Map output bytes=14451444
```

Fig. 4. MapReduce Shell Script a


Activities
Terminal
Thu 17:33
hduser@shikhar-VirtualBox: ~/OlympicFiltering/output
File Edit View Search Terminal Help

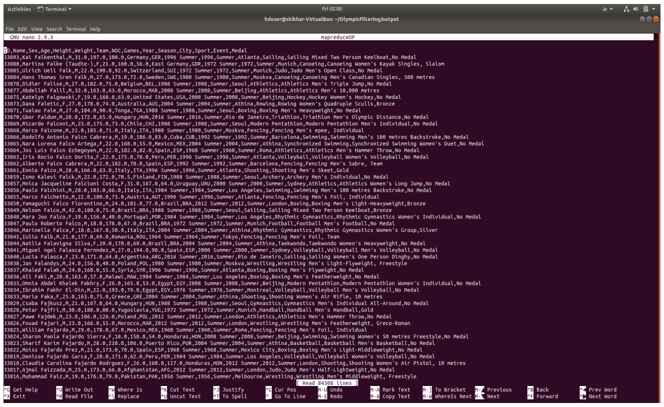
```
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=178703
                Total time spent by all reduces in occupied slots (ms)=17524
                Total time spent by all map tasks (ms)=178703
                Total time spent by all reduce tasks (ms)=17524
                Total vcore-milliseconds taken by all map tasks=178703
                Total vcore-milliseconds taken by all reduce tasks=17524
                Total megabyte-milliseconds taken by all map tasks=182991872
                Total megabyte-milliseconds taken by all reduce tasks=17944576
        Map-Reduce Framework
                Map input records=99088
                Map output records=99088
                Map output bytes=14451444
                Map output materialized bytes=14730486
                Input split bytes=480
                Combine input records=0
                Combine output records=0
                Reduce input groups=49
                Reduce shuffle bytes=14730486
                Reduce input records=99088
                Reduce output records=84307
                Spilled Records=198176
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=2939
                CPU time spent (ms)=11780
                Physical memory (bytes) snapshot=989679616
                Virtual memory (bytes) snapshot=9645404160
                Total committed heap usage (bytes)=700776448
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=13760130
        File Output Format Counters
                Bytes Written=11650844
hduser@shikhar-VirtualBox:~/OlympicFiltering$ ls -l
total 16
-rw-r--r-- 1 hduser hadoopgroup 4697 Apr 25 17:15 Filtering.jar
-rwxr-xr-x 1 hduser hadoopgroup  948 Apr 25 14:37 filtering.sh
drwxr-xr-x 2 hduser hadoopgroup 4096 Apr 25 17:22 output
hduser@shikhar-VirtualBox:~/OlympicFiltering$ cd output/
hduser@shikhar-VirtualBox:~/OlympicFiltering/output$ sudo nano part-r-00000
hduser@shikhar-VirtualBox:~/OlympicFiltering/output$ 
```

Fig. 5. MapReduce Shell Script b


Activities
Terminal
Thu 17:37
hduser@shikhar-VirtualBox: ~/OlympicFiltering/output
File Edit View Search Terminal Help

```
  GNU nano 2.9.3                                      part-r-00000

        33893,Kai Falkenthal,M,31.0,197.0,100.0,Germany,GER,1996 Summer,1996,Summer,Atlanta,Sailing,Sailing Mixed Two Person Keelboat,No Medal
        33888,Martina Falke (Taudte-),F,21.0,160.0,56.0,East Germany,GDR,1972 Summer,1972,Summer,Munich,Canoeing,Canoeing Women's Kayak Singles, Slalom
        33885,Ulrich Ueli Falk,M,22.0,190.0,92.0,Switzerland,SUI,1972 Summer,1972,Summer,Munich,Judo,Judo Men's Open Class,No Medal
        33884,Hans Thomas Sren Falk,M,27.0,173.0,72.0,Sweden,SWE,1980 Summer,1980,Summer,Moskva,Canoeing,Canoeing Men's Canadian Singles, 500 metres
        33878,Didier Falise,M,27.0,182.0,75.0,Belgium,BEL,1988 Summer,1988,Summer,Seoul,Athletics,Athletics Men's Triple Jump,No Medal
        33877,Abdellah Falil,M,32.0,163.0,63.0,Morocco,MAR,2008 Summer,2008,Summer,Beijing,Athletics,Athletics Men's 10,000 metres
        33875,Katelyn Falgowski,F,19.0,168.0,63.0,United States,USA,2008 Summer,2008,Summer,Beijing,Hockey,Hockey Women's Hockey,No Medal
        33873,Dana Faletic,F,27.0,178.0,74.0,Australia,AUS,2004 Summer,2004,Summer,Athina,Rowing,Rowing Women's Quadruple Sculls,Bronze
        33871,Tualau Fale,M,27.0,184.0,90.0,Tonga,TGA,1988 Summer,1988,Summer,Seoul,Boxing,Boxing Men's Heavyweight,No Medal
        33870,Gbor Faldum,M,28.0,172.0,65.0,Hungary,HUN,2016 Summer,2016,Summer,Rio de Janeiro,Triathlon,Triathlon Men's Olympic Distance,No Medal
        33869,Ricardo Falconi,M,25.0,175.0,73.0,Chile,CHI,1988 Summer,1988,Summer,Seoul,Modern Pentathlon,Modern Pentathlon Men's Individual,No Medal
        33868,Marco Falcone,M,21.0,185.0,71.0,Italy,ITA,1980 Summer,1980,Summer,Moskva,Fencing,Fencing Men's epee, Individual
        33866,Rodolfo Antonio Falcn Cabrera,M,19.0,186.0,83.0,Cuba,CUB,1992 Summer,1992,Summer,Barcelona,Swimming,Swimming Men's 100 metres Backstroke,No Medal
        33865,Nara Lorena Falcn Artega,F,22.0,168.0,55.0,Mexico,MEX,2004 Summer,2004,Summer,Athina,Synchronized Swimming,Synchronized Swimming Women's Duet,No Medal
        33864,Jos Luis Falcn Echegoyen,M,22.0,182.0,82.0,Spain,ESP,1960 Summer,1960,Summer,Roma,Athletics,Athletics Men's Hammer Throw,No Medal
        33863,Iris Rocio Falcn Dorita,F,22.0,175.0,78.0,Peru,PER,1996 Summer,1996,Summer,Atlanta,Volleyball,Volleyball Women's Volleyball,No Medal
        33862,Alberto Falcn Cabrera,M,22.0,182.0,78.0,Spain,ESP,1992 Summer,1992,Summer,Barcelona,Fencing,Fencing Men's Sabre, Team
        33861,Ennio Falco,M,28.0,166.0,63.0,Italy,ITA,1996 Summer,1996,Summer,Atlanta,Shooting,Shooting Men's Skeet,Gold
        33859,Ismo Kalevi Falck,M,22.0,172.0,70.5,Finland,FIN,1988 Summer,1988,Summer,Seoul,Archery,Archery Men's Individual,No Medal
        33857,Mnica Jacqueline Falcioni Costa,F,31.0,167.0,64.0,Uruguay,URU,2000 Summer,2000,Summer,Sydney,Athletics,Athletics Women's Long Jump,No Medal
        33856,Paolo Falchini,M,20.0,183.0,66.0,Italy,ITA,1984 Summer,1984,Summer,Los Angeles,Swimming,Swimming Men's 100 metres Backstroke,No Medal
        33855,Marco Falchetto,M,22.0,180.0,73.0,Austria,AUT,1996 Summer,1996,Summer,Atlanta,Fencing,Fencing Men's Foil, Individual
        33850,Yamaguchi Falco Florentino,M,24.0,181.0,77.0,Brazil,BRA,2012 Summer,2012,Summer,London,Boxing,Boxing Men's Light-Heavyweight,Bronze
        33849,Nelson Falco,M,42.0,180.0,75.0,Brazil,BRA,1988 Summer,1988,Summer,Seoul,Sailing,Sailing Mixed Two Person Keelboat,Bronze
        33848,Mara Joo Falco,F,19.0,156.0,40.0,Portugal,POR,1984 Summer,1984,Summer,Los Angeles,Rhythmic Gymnastics,Rhythmic Gymnastics Women's Individual,No Medal
        33847,Paulo Roberto Falco,M,18.0,178.0,67.0,Brazil,BRA,1972 Summer,1972,Summer,Munich,Football,Football Men's Football,No Medal
        33846,Marinella Falca,F,18.0,167.0,50.0,Italy,ITA,2004 Summer,2004,Summer,Athina,Rhythmic Gymnastics,Rhythmic Gymnastics Women's Group,Silver
        33845,Iuliu Falb,M,21.0,177.0,69.0,Romania,ROU,1964 Summer,1964,Summer,Tokyo,Fencing,Fencing Men's Foil, Team
        33844,Natlia Falavigna Silva,F,20.0,178.0,69.0,Brazil,BRA,2004 Summer,2004,Summer,Athina,Taekwondo,Taekwondo Women's Heavyweight,No Medal
        33841,Miguel ngel Falasca Fernndez,M,27.0,194.0,90.0,Spain,ESP,2000 Summer,2000,Summer,Sydney,Volleyball,Volleyball Men's Volleyball,No Medal
        33840,Lucia Falasca,F,23.0,171.0,64.0,Argentina,ARG,2016 Summer,2016,Summer,Rio de Janeiro,Sailing,Sailing Women's One Person Dinghy,No Medal
        33838,Jan Falandys,M,24.0,156.0,48.0,Poland,POL,1980 Summer,1980,Summer,Moskva,Wrestling,Wrestling Men's Light-Flyweight, Freestyle
        33837,Khaled Falah,M,24.0,160.0,51.0,Syria,SYR,1996 Summer,1996,Summer,Atlanta,Boxing,Boxing Men's Flyweight,No Medal
        33836,Ali Faki,M,20.0,163.0,57.0,Malawi,MAW,1984 Summer,1984,Summer,Los Angeles,Boxing,Boxing Men's Featherweight,No Medal
        33835,Omnia Abdel Khalek Fakhry,F,26.0,165.0,53.0,Egypt,EGY,2008 Summer,2008,Summer,Beijing,Modern Pentathlon,Modern Pentathlon Women's Individual,No Medal
        33834,Ibrahim Fakhr El-Din,M,22.0,193.0,79.0,Egypt,EGY,1976 Summer,1976,Summer,Montreal,Volleyball,Volleyball Men's Volleyball,No Medal
        33833,Maria Faka,F,25.0,163.0,75.0,Greece,GRE,2004 Summer,2004,Summer,Athina,Shooting,Shooting Women's Air Rifle, 10 metres
        33829,Csaba Fajkusz,M,21.0,167.0,64.0,Hungary,HUN,1988 Summer,1988,Summer,Seoul,Gymnastics,Gymnastics Men's Individual All-Around,No Medal
        33828,Petar Fajfri,M,30.0,180.0,80.0,Yugoslavia,YUG,1972 Summer,1972,Summer,Munich,Handball,Handball Men's Handball,Gold
        33827,Pawe Fajdek,M,23.0,186.0,126.0,Poland,POL,2012 Summer,2012,Summer,London,Athletics,Athletics Men's Hammer Throw,No Medal
        33826,Fouad Fajari,M,23.0,166.0,55.0,Morocco,MAR,2012 Summer,2012,Summer,London,Wrestling,Wrestling Men's Featherweight, Greco-Roman
        33825,William Fajardo,M,29.0,178.0,67.0,Mexico,MEX,1960 Summer,1960,Summer,Roma,Fencing,Fencing Men's Foil, Individual
        33824,Sharon Paola Fajardo Sierra,F,18.0,158.0,54.0,Honduras,HON,2008 Summer,2008,Summer,Beijing,Swimming,Swimming Women's 50 metres Freestyle,No Medal
        33823,Sharif Karim Fajardo,M,28.0,210.0,106.0,Puerto Rico,PUR,2004 Summer,2004,Summer,Athina,Basketball,Basketball Men's Basketball,No Medal
        33822,Moiss Fajardo Prez,M,21.0,173.0,70.0,Spain,ESP,1968 Summer,1968,Summer,Mexico City,Boxing,Boxing Men's Light-Middleweight,No Medal
        33819,Denisse Fajardo Garca,F,20.0,171.0,62.0,Peru,PER,1984 Summer,1984,Summer,Los Angeles,Volleyball,Volleyball Women's Volleyball,No Medal
        33818,Claudia Carolina Fajardo Rodrguez,F,26.0,160.0,117.0,Honduras,HON,2012 Summer,2012,Summer,London,Shooting,Shooting Women's Air Pistol, 10 metres
        33817,Ajmal Faizzada,M,25.0,173.0,66.0,Afghanistan,AFG,2012 Summer,2012,Summer,London,Judo,Judo Men's Half-Lightweight,No Medal
        33816,Muhammad Faiz,M,19.0,176.0,79.0,Pakistan,PAK,1956 Summer,1956,Summer,Melbourne,Wrestling,Wrestling Men's Middleweight, Freestyle
        33815,Didier Faivre-Pierret,M,27.0,182.0,70.0,France,FRA,1992 Summer,1992,Summer,Barcelona,Cycling,Cycling Men's 100 kilometres Team Time Trial,Bronze
                                                  [ Read 84307 lines ]
^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo      M-A Mark Text   M-] To Bracket   M-▲ Previous   ^B Back      ^◀ Prev Word
^X Exit       ^R Read File  ^\ Replace    ^U Uncut Text ^T To Spell   ^_ Go To Line M-E Redo      M-6 Copy Text   M-W WhereIs Next M-▼ Next       ^F Forward   ^▶ Next Word
```

Fig. 6. MapReduce Output


Activities
Terminal
Fri 02:00
ie
hduser@shikhar-VirtualBox: ~/OlympicFiltering/output
File Edit View Search Terminal Help

```
  GNU nano 2.9.3                                      mapreduceOP

ID,Name,Sex,Age,Height,Weight,Team,NOC,Games,Year,Season,City,Sport,Event,Medal
33893,Kai Falkenthal,M,31.0,197.0,100.0,Germany,GER,1996 Summer,1996,Summer,Atlanta,Sailing,Sailing Mixed Two Person Keelboat,No Medal
33888,Martina Falke (Taudte-),F,21.0,160.0,56.0,East Germany,GDR,1972 Summer,1972,Summer,Munich,Canoeing,Canoeing Women's Kayak Singles, Slalom
33885,Ulrich Ueli Falk,M,22.0,190.0,92.0,Switzerland,SUI,1972 Summer,1972,Summer,Munich,Judo,Judo Men's Open Class,No Medal
33884,Hans Thomas Sren Falk,M,27.0,173.0,72.0,Sweden,SWE,1980 Summer,1980,Summer,Moskva,Canoeing,Canoeing Men's Canadian Singles, 500 metres
33878,Didier Falise,M,27.0,182.0,75.0,Belgium,BEL,1988 Summer,1988,Summer,Seoul,Athletics,Athletics Men's Triple Jump,No Medal
33877,Abdellah Falil,M,32.0,163.0,63.0,Morocco,MAR,2008 Summer,2008,Summer,Beijing,Athletics,Athletics Men's 10,000 metres
33875,Katelyn Falgowski,F,19.0,168.0,63.0,United States,USA,2008 Summer,2008,Summer,Beijing,Hockey,Hockey Women's Hockey,No Medal
33873,Dana Faletic,F,27.0,178.0,74.0,Australia,AUS,2004 Summer,2004,Summer,Athina,Rowing,Rowing Women's Quadruple Sculls,Bronze
33871,Tualau Fale,M,27.0,184.0,90.0,Tonga,TGA,1988 Summer,1988,Summer,Seoul,Boxing,Boxing Men's Heavyweight,No Medal
33870,Gbor Faldum,M,28.0,172.0,65.0,Hungary,HUN,2016 Summer,2016,Summer,Rio de Janeiro,Triathlon,Triathlon Men's Olympic Distance,No Medal
33869,Ricardo Falconi,M,25.0,175.0,73.0,Chile,CHI,1988 Summer,1988,Summer,Seoul,Modern Pentathlon,Modern Pentathlon Men's Individual,No Medal
33868,Marco Falcone,M,21.0,185.0,71.0,Italy,ITA,1980 Summer,1980,Summer,Moskva,Fencing,Fencing Men's epee, Individual
33866,Rodolfo Antonio Falcn Cabrera,M,19.0,186.0,83.0,Cuba,CUB,1992 Summer,1992,Summer,Barcelona,Swimming,Swimming Men's 100 metres Backstroke,No Medal
33865,Nara Lorena Falcn Artega,F,22.0,168.0,55.0,Mexico,MEX,2004 Summer,2004,Summer,Athina,Synchronized Swimming,Synchronized Swimming Women's Duet,No Medal
33864,Jos Luis Falcn Echegoyen,M,22.0,182.0,82.0,Spain,ESP,1960 Summer,1960,Summer,Roma,Athletics,Athletics Men's Hammer Throw,No Medal
33863,Iris Rocio Falcn Dorita,F,22.0,175.0,78.0,Peru,PER,1996 Summer,1996,Summer,Atlanta,Volleyball,Volleyball Women's Volleyball,No Medal
33862,Alberto Falcn Cabrera,M,22.0,182.0,78.0,Spain,ESP,1992 Summer,1992,Summer,Barcelona,Fencing,Fencing Men's Sabre, Team
33861,Ennio Falco,M,28.0,166.0,63.0,Italy,ITA,1996 Summer,1996,Summer,Atlanta,Shooting,Shooting Men's Skeet,Gold
33859,Ismo Kalevi Falck,M,22.0,172.0,70.5,Finland,FIN,1988 Summer,1988,Summer,Seoul,Archery,Archery Men's Individual,No Medal
33857,Mnica Jacqueline Falcioni Costa,F,31.0,167.0,64.0,Uruguay,URU,2000 Summer,2000,Summer,Sydney,Athletics,Athletics Women's Long Jump,No Medal
33856,Paolo Falchini,M,20.0,183.0,66.0,Italy,ITA,1984 Summer,1984,Summer,Los Angeles,Swimming,Swimming Men's 100 metres Backstroke,No Medal
33855,Marco Falchetto,M,22.0,180.0,73.0,Austria,AUT,1996 Summer,1996,Summer,Atlanta,Fencing,Fencing Men's Foil, Individual
33850,Yamaguchi Falco Florentino,M,24.0,181.0,77.0,Brazil,BRA,2012 Summer,2012,Summer,London,Boxing,Boxing Men's Light-Heavyweight,Bronze
33849,Nelson Falco,M,42.0,180.0,75.0,Brazil,BRA,1988 Summer,1988,Summer,Seoul,Sailing,Sailing Mixed Two Person Keelboat,Bronze
33848,Mara Joo Falco,F,19.0,156.0,40.0,Portugal,POR,1984 Summer,1984,Summer,Los Angeles,Rhythmic Gymnastics,Rhythmic Gymnastics Women's Individual,No Medal
33847,Paulo Roberto Falco,M,18.0,178.0,67.0,Brazil,BRA,1972 Summer,1972,Summer,Munich,Football,Football Men's Football,No Medal
33846,Marinella Falca,F,18.0,167.0,50.0,Italy,ITA,2004 Summer,2004,Summer,Athina,Rhythmic Gymnastics,Rhythmic Gymnastics Women's Group,Silver
33845,Iuliu Falb,M,21.0,177.0,69.0,Romania,ROU,1964 Summer,1964,Summer,Tokyo,Fencing,Fencing Men's Foil, Team
33844,Natlia Falavigna Silva,F,20.0,178.0,69.0,Brazil,BRA,2004 Summer,2004,Summer,Athina,Taekwondo,Taekwondo Women's Heavyweight,No Medal
33841,Miguel ngel Falasca Fernndez,M,27.0,194.0,90.0,Spain,ESP,2000 Summer,2000,Summer,Sydney,Volleyball,Volleyball Men's Volleyball,No Medal
33840,Lucia Falasca,F,23.0,171.0,64.0,Argentina,ARG,2016 Summer,2016,Summer,Rio de Janeiro,Sailing,Sailing Women's One Person Dinghy,No Medal
33838,Jan Falandys,M,24.0,156.0,48.0,Poland,POL,1980 Summer,1980,Summer,Moskva,Wrestling,Wrestling Men's Light-Flyweight, Freestyle
33837,Khaled Falah,M,24.0,160.0,51.0,Syria,SYR,1996 Summer,1996,Summer,Atlanta,Boxing,Boxing Men's Flyweight,No Medal
33836,Ali Faki,M,20.0,163.0,57.0,Malawi,MAW,1984 Summer,1984,Summer,Los Angeles,Boxing,Boxing Men's Featherweight,No Medal
33835,Omnia Abdel Khalek Fakhry,F,26.0,165.0,53.0,Egypt,EGY,2008 Summer,2008,Summer,Beijing,Modern Pentathlon,Modern Pentathlon Women's Individual,No Medal
33834,Ibrahim Fakhr El-Din,M,22.0,193.0,79.0,Egypt,EGY,1976 Summer,1976,Summer,Montreal,Volleyball,Volleyball Men's Volleyball,No Medal
33833,Maria Faka,F,25.0,163.0,75.0,Greece,GRE,2004 Summer,2004,Summer,Athina,Shooting,Shooting Women's Air Rifle, 10 metres
33829,Csaba Fajkusz,M,21.0,167.0,64.0,Hungary,HUN,1988 Summer,1988,Summer,Seoul,Gymnastics,Gymnastics Men's Individual All-Around,No Medal
33828,Petar Fajfri,M,30.0,180.0,80.0,Yugoslavia,YUG,1972 Summer,1972,Summer,Munich,Handball,Handball Men's Handball,Gold
33827,Pawe Fajdek,M,23.0,186.0,126.0,Poland,POL,2012 Summer,2012,Summer,London,Athletics,Athletics Men's Hammer Throw,No Medal
33826,Fouad Fajari,M,23.0,166.0,55.0,Morocco,MAR,2012 Summer,2012,Summer,London,Wrestling,Wrestling Men's Featherweight, Greco-Roman
33825,William Fajardo,M,29.0,178.0,67.0,Mexico,MEX,1960 Summer,1960,Summer,Roma,Fencing,Fencing Men's Foil, Individual
33824,Sharon Paola Fajardo Sierra,F,18.0,158.0,54.0,Honduras,HON,2008 Summer,2008,Summer,Beijing,Swimming,Swimming Women's 50 metres Freestyle,No Medal
33823,Sharif Karim Fajardo,M,28.0,210.0,106.0,Puerto Rico,PUR,2004 Summer,2004,Summer,Athina,Basketball,Basketball Men's Basketball,No Medal
33822,Moiss Fajardo Prez,M,21.0,173.0,70.0,Spain,ESP,1968 Summer,1968,Summer,Mexico City,Boxing,Boxing Men's Light-Middleweight,No Medal
33819,Denisse Fajardo Garca,F,20.0,171.0,62.0,Peru,PER,1984 Summer,1984,Summer,Los Angeles,Volleyball,Volleyball Women's Volleyball,No Medal
33818,Claudia Carolina Fajardo Rodrguez,F,26.0,160.0,117.0,Honduras,HON,2012 Summer,2012,Summer,London,Shooting,Shooting Women's Air Pistol, 10 metres
33817,Ajmal Faizzada,M,25.0,173.0,66.0,Afghanistan,AFG,2012 Summer,2012,Summer,London,Judo,Judo Men's Half-Lightweight,No Medal
33816,Muhammad Faiz,M,19.0,176.0,79.0,Pakistan,PAK,1956 Summer,1956,Summer,Melbourne,Wrestling,Wrestling Men's Middleweight, Freestyle
                                                  [ Read 84308 lines ]
^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo      M-A Mark Text   M-] To Bracket   M-▲ Previous   ^B Back      ^◀ Prev Word
^X Exit       ^R Read File  ^\ Replace    ^U Uncut Text ^T To Spell   ^_ Go To Line M-E Redo      M-6 Copy Text   M-W WhereIs Next M-▼ Next       ^F Forward   ^▶ Next Word
```

Fig. 7. Processed MapReduce Output

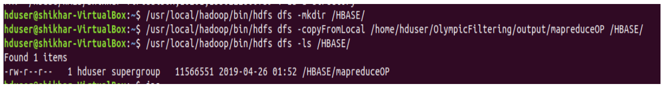
hduser in 'OlympicFiltering' directory. A shell script was then written to execute MapReduce jar file on Olympic data which was sqooped from MySQL to HDFS. Jar file and the shell script written are present in the compressed folder uploaded. Fig. 4 and Fig. 5 show the execution of MapReduce task on HDFS Olympic data.

Step 4: Output from MapReduce task is shown in Fig. 6. The output data from MapReduce task was not looking good because of the space before each ID and the headers were also not present. The space before IDs were because a blank field was included in the Reducer class while writing the output after filtering. Due to this, output file from MapReduce task was then processed to make it look like a neat output. Fig. 7 shows the neat file after processing output from MapReduce.

Step 5: This Step involves putting organized MapReduce output data in HDFS to further load it in HBase. Script used to load MapReduce output data in HDFS is shown in Fig. 8. A separate directory was created named 'HBASE' to store MapReduce output data in HDFS.

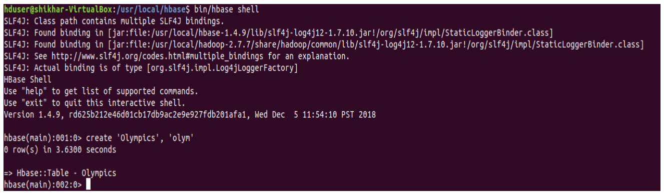
Step 6: After putting MapReduce output data in HDFS, HBase was then started and launched. A new table was created in HBase to store the output from MapReduce. Fig. 9 shows screenshot for creating a new table in HBase and Fig. 10 then shows data loading in HBase. The HBase table data is also shown in Fig. 11

Step 7: Now, with an aim to pull MapReduce output data in Hive from MySQL using Sqoop for further analysis, a new MySQL table was created named 'summer_olympics'. Fig. 12 shows the creation of a new table and some sample records from that table. This table holds data for only summer Olympics as data put into this table is coming from MapRe-



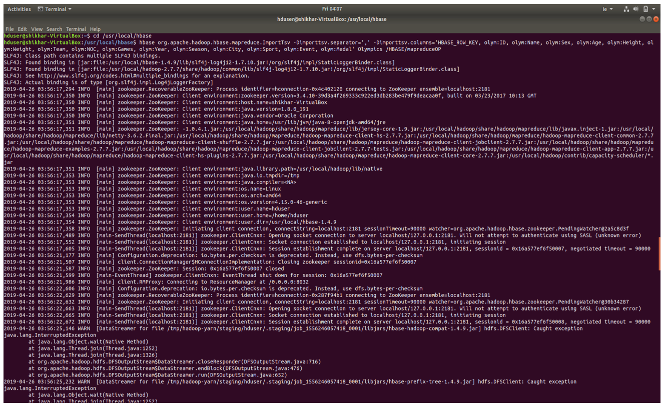
```
hduser@shikhar-VirtualBox:~$ /usr/local/hadoop/bin/hdfs dfs -mkdir /HBASE/
hduser@shikhar-VirtualBox:~$ /usr/local/hadoop/bin/hdfs dfs -copyFromLocal /home/hduser/OlympicFiltering/output/mapreduceOP /HBASE/
hduser@shikhar-VirtualBox:~$ /usr/local/hadoop/bin/hdfs dfs -ls /HBASE/
Found 1 items
-rw-r--r--   1 hduser supergroup   11566551 2019-04-26 01:52 /HBASE/mapreduceOP
```

Fig. 8. MapReduce Output to HDFS for HBASE



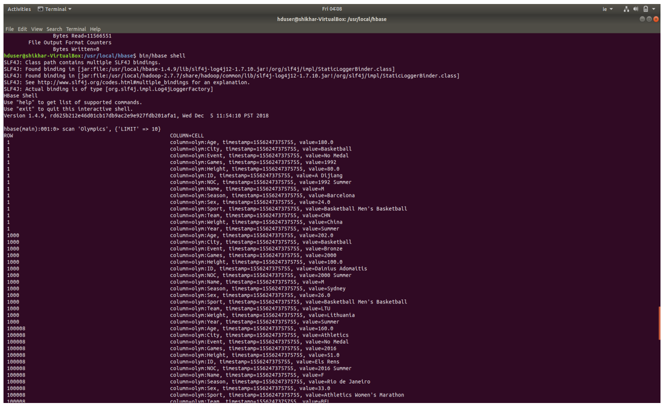
```
hduser@shikhar-VirtualBox:/usr/local/hbase$ bin/hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.9, rd625b212e46d01cb17db9ac2e9e927fdb201afa1, Wed Dec  5 11:54:10 PST 2018

hbase(main):001:0> create 'Olympics', 'olym'
0 row(s) in 3.6300 seconds

=> Hbase::Table - Olympics
hbase(main):002:0> 
```

Fig. 9. Table Created in HBase


Activities
Terminal
Fri 04:07
ie
hduser@shikhar-VirtualBox: /usr/local/hbase
File Edit View Search Terminal Help

```
hduser@shikhar-VirtualBox:~$ cd /usr/local/hbase
hduser@shikhar-VirtualBox:/usr/local/hbase$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY, olym:ID, olym:Name, olym:Sex, olym:Age, olym:Height, ol
ym:Weight, olym:Team, olym:NOC, olym:Games, olym:Year, olym:Season, olym:City, olym:Sport, olym:Event, olym:Medal' Olympics /HBASE/mapreduceOP
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2019-04-26 03:56:17,294 INFO  [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x4c402120 connecting to ZooKeeper ensemble=localhost:2181
2019-04-26 03:56:17,350 INFO  [main] zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.10-39d3a4f269333c922ed3db283be479f9deacaa0f, built on 03/23/2017 10:13 GMT
2019-04-26 03:56:17,350 INFO  [main] zookeeper.ZooKeeper: Client environment:host.name=shikhar-VirtualBox
2019-04-26 03:56:17,350 INFO  [main] zookeeper.ZooKeeper: Client environment:java.version=1.8.0_191
2019-04-26 03:56:17,350 INFO  [main] zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation
2019-04-26 03:56:17,351 INFO  [main] zookeeper.ZooKeeper: Client environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre
2019-04-26 03:56:17,351 INFO  [main] zookeeper.ZooKeeper: -1.0.4.1.jar:/usr/local/hadoop/share/hadoop/mapreduce/lib/jersey-core-1.9.jar:/usr/local/hadoop/share/hadoop/mapreduce/lib/javax.inject-1.jar:/usr/local/
hadoop/share/hadoop/mapreduce/lib/netty-3.6.2.Final.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-hs-2.7.7.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-common-2.7.7
.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-shuffle-2.7.7.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.7.jar:/usr/local/hadoop/share/hadoop/mapredu
ce/hadoop-mapreduce-examples-2.7.7.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.7-tests.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-app-2.7.7.jar:/u
sr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-hs-plugins-2.7.7.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.7.7.jar:/usr/local/hadoop/contrib/capacity-scheduler/*.
jar
2019-04-26 03:56:17,351 INFO  [main] zookeeper.ZooKeeper: Client environment:java.library.path=/usr/local/hadoop/lib/native
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:os.name=Linux
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:os.arch=amd64
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:os.version=4.15.0-46-generic
2019-04-26 03:56:17,353 INFO  [main] zookeeper.ZooKeeper: Client environment:user.name=hduser
2019-04-26 03:56:17,354 INFO  [main] zookeeper.ZooKeeper: Client environment:user.home=/home/hduser
2019-04-26 03:56:17,354 INFO  [main] zookeeper.ZooKeeper: Client environment:user.dir=/usr/local/hbase-1.4.9
2019-04-26 03:56:17,358 INFO  [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.PendingWatcher@2a5c8d3f
2019-04-26 03:56:17,489 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate using SASL (unknown error)
2019-04-26 03:56:17,552 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
2019-04-26 03:56:17,605 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x16a577ef6f50007, negotiated timeout = 90000
2019-04-26 03:56:21,177 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-26 03:56:21,507 INFO  [main] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x16a577ef6f50007
2019-04-26 03:56:21,587 INFO  [main] zookeeper.ZooKeeper: Session: 0x16a577ef6f50007 closed
2019-04-26 03:56:21,599 INFO  [main-EventThread] zookeeper.ClientCnxn: EventThread shut down for session: 0x16a577ef6f50007
2019-04-26 03:56:21,986 INFO  [main] client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-04-26 03:56:22,606 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-26 03:56:22,629 INFO  [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x287f94b1 connecting to ZooKeeper ensemble=localhost:2181
2019-04-26 03:56:22,632 INFO  [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.PendingWatcher@30b34287
2019-04-26 03:56:22,663 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate using SASL (unknown error)
2019-04-26 03:56:22,665 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
2019-04-26 03:56:22,672 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x16a577ef6f50008, negotiated timeout = 90000
2019-04-26 03:56:25,146 WARN  [DataStreamer for file /tmp/hadoop-yarn/staging/hduser/.staging/job_1556246057418_0001/libjars/hbase-hadoop-compat-1.4.9.jar] hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:716)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:476)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:652)
2019-04-26 03:56:25,232 WARN  [DataStreamer for file /tmp/hadoop-yarn/staging/hduser/.staging/job_1556246057418_0001/libjars/hbase-prefix-tree-1.4.9.jar] hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
```

Fig. 10. Importing MapReduce Output Data in HBASE

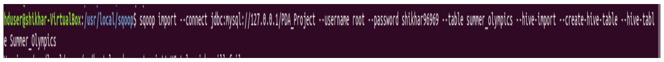

Activities
Terminal
Fri 04:08
ie
hduser@shikhar-VirtualBox: /usr/local/hbase
File Edit View Search Terminal Help

```
                Bytes Read=11566551
        File Output Format Counters
                Bytes Written=0
hduser@shikhar-VirtualBox:/usr/local/hbase$ bin/hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.9, rd625b212e46d01cb17db9ac2e9e927fdb201afa1, Wed Dec  5 11:54:10 PST 2018

hbase(main):001:0> scan 'Olympics', {'LIMIT' => 10}
ROW                                                  COLUMN+CELL
 1                                                   column=olym:Age, timestamp=1556247375755, value=180.0
 1                                                   column=olym:City, timestamp=1556247375755, value=Basketball
 1                                                   column=olym:Event, timestamp=1556247375755, value=No Medal
 1                                                   column=olym:Games, timestamp=1556247375755, value=1992
 1                                                   column=olym:Height, timestamp=1556247375755, value=80.0
 1                                                   column=olym:ID, timestamp=1556247375755, value=A Dijiang
 1                                                   column=olym:NOC, timestamp=1556247375755, value=1992 Summer
 1                                                   column=olym:Name, timestamp=1556247375755, value=M
 1                                                   column=olym:Season, timestamp=1556247375755, value=Barcelona
 1                                                   column=olym:Sex, timestamp=1556247375755, value=24.0
 1                                                   column=olym:Sport, timestamp=1556247375755, value=Basketball Men's Basketball
 1                                                   column=olym:Team, timestamp=1556247375755, value=CHN
 1                                                   column=olym:Weight, timestamp=1556247375755, value=China
 1                                                   column=olym:Year, timestamp=1556247375755, value=Summer
 1000                                                column=olym:Age, timestamp=1556247375755, value=202.0
 1000                                                column=olym:City, timestamp=1556247375755, value=Basketball
 1000                                                column=olym:Event, timestamp=1556247375755, value=Bronze
 1000                                                column=olym:Games, timestamp=1556247375755, value=2000
 1000                                                column=olym:Height, timestamp=1556247375755, value=100.0
 1000                                                column=olym:ID, timestamp=1556247375755, value=Dainius Adomaitis
 1000                                                column=olym:NOC, timestamp=1556247375755, value=2000 Summer
 1000                                                column=olym:Name, timestamp=1556247375755, value=M
 1000                                                column=olym:Season, timestamp=1556247375755, value=Sydney
 1000                                                column=olym:Sex, timestamp=1556247375755, value=26.0
 1000                                                column=olym:Sport, timestamp=1556247375755, value=Basketball Men's Basketball
 1000                                                column=olym:Team, timestamp=1556247375755, value=LTU
 1000                                                column=olym:Weight, timestamp=1556247375755, value=Lithuania
 1000                                                column=olym:Year, timestamp=1556247375755, value=Summer
 100008                                              column=olym:Age, timestamp=1556247375755, value=160.0
 100008                                              column=olym:City, timestamp=1556247375755, value=Athletics
 100008                                              column=olym:Event, timestamp=1556247375755, value=No Medal
 100008                                              column=olym:Games, timestamp=1556247375755, value=2016
 100008                                              column=olym:Height, timestamp=1556247375755, value=51.0
 100008                                              column=olym:ID, timestamp=1556247375755, value=Els Rens
 100008                                              column=olym:NOC, timestamp=1556247375755, value=2016 Summer
 100008                                              column=olym:Name, timestamp=1556247375755, value=F
 100008                                              column=olym:Season, timestamp=1556247375755, value=Rio de Janeiro
 100008                                              column=olym:Sex, timestamp=1556247375755, value=33.0
 100008                                              column=olym:Sport, timestamp=1556247375755, value=Athletics Women's Marathon
 100008                                              column=olym:Team, timestamp=1556247375755, value=BEL
```

Fig. 11. HBase Table Data

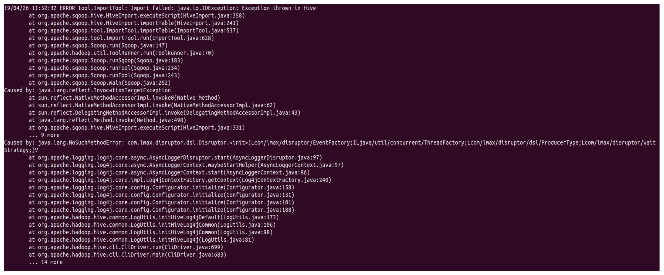
duce output.

Step 8: This involved importing data from newly created summer Olympics table from MySQL to Hive. But, the command was giving unusual errors when the script was executed. Fig. 13 show the first command which threw error when trying to sqoop data from MySQL to Hive. This command was supposed to create a new table in Hive and then put MySQL data in that table.



```
hduser@shikhar-VirtualBox:/usr/local/sqoop$ sqoop import --connect jdbc:mysql://127.0.0.1/PDA_Project --username root --password shikhar96969 --table summer_olympics --hive-import --create-hive-table --hive-tabl
e Summer_Olympics
```

Fig. 13. First Sqoop Error Command



```
19/04/26 11:52:32 ERROR tool.ImportTool: Import failed: java.io.IOException: Exception thrown in Hive
        at org.apache.sqoop.hive.HiveImport.executeScript(HiveImport.java:358)
        at org.apache.sqoop.hive.HiveImport.importTable(HiveImport.java:241)
        at org.apache.sqoop.tool.ImportTool.importTable(ImportTool.java:537)
        at org.apache.sqoop.tool.ImportTool.run(ImportTool.java:628)
        at org.apache.sqoop.Sqoop.run(Sqoop.java:147)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
        at org.apache.sqoop.Sqoop.runSqoop(Sqoop.java:183)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:234)
        at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
        at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
Caused by: java.lang.reflect.InvocationTargetException
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.sqoop.hive.HiveImport.executeScript(HiveImport.java:331)
        ... 9 more
Caused by: java.lang.NoSuchMethodError: com.lmax.disruptor.dsl.Disruptor.<init>(Lcom/lmax/disruptor/EventFactory;ILjava/util/concurrent/ThreadFactory;Lcom/lmax/disruptor/dsl/ProducerType;Lcom/lmax/disruptor/Wait
Strategy;)V
        at org.apache.logging.log4j.core.async.AsyncLoggerDisruptor.start(AsyncLoggerDisruptor.java:97)
        at org.apache.logging.log4j.core.async.AsyncLoggerContext.maybeStartHelper(AsyncLoggerContext.java:97)
        at org.apache.logging.log4j.core.async.AsyncLoggerContext.start(AsyncLoggerContext.java:86)
        at org.apache.logging.log4j.core.impl.Log4jContextFactory.getContext(Log4jContextFactory.java:240)
        at org.apache.logging.log4j.core.config.Configurator.initialize(Configurator.java:158)
        at org.apache.logging.log4j.core.config.Configurator.initialize(Configurator.java:131)
        at org.apache.logging.log4j.core.config.Configurator.initialize(Configurator.java:101)
        at org.apache.logging.log4j.core.config.Configurator.initialize(Configurator.java:188)
        at org.apache.hadoop.hive.common.LogUtils.initHiveLog4jDefault(LogUtils.java:173)
        at org.apache.hadoop.hive.common.LogUtils.initHiveLog4jCommon(LogUtils.java:106)
        at org.apache.hadoop.hive.common.LogUtils.initHiveLog4jCommon(LogUtils.java:98)
        at org.apache.hadoop.hive.common.LogUtils.initHiveLog4j(LogUtils.java:81)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:699)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        ... 14 more
```

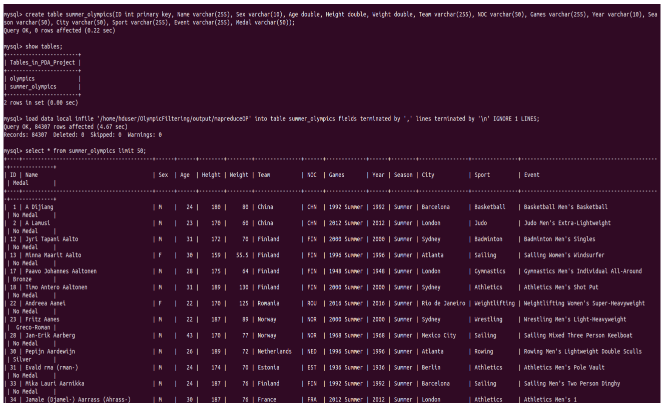
Fig. 14. First Sqoop Command Error

It gave an error in executeScript while importing data from MySQL as shown in Fig. 14. With an assumption that error is caused due to primary key which was present in MySQL table being imported, a new table was created in Hive manually and then Sqooping was tried again, but this time create hive table was removed from the script. This also didn't work out and this command gave new error of HDFS directory already exists as shown in Fig. 15.

The HDFS directory was then manually deleted as shown in Fig. 16 and command was executed again. Unfortunately, this also didn't work out and gave same error of executeScript failure.

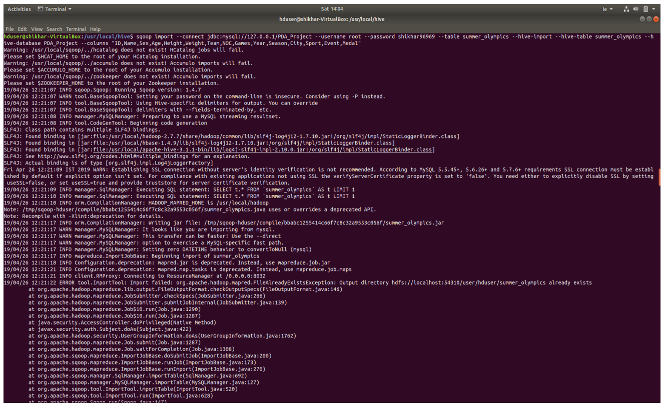
Step 9: Due to the errors faced while sqooping data from MySQL to Hive, data was manually inserted in Hive table (shown in Fig. 17) and queries were then executed on this table.

Step 10: In this step four queries were written to analyze data loaded in Hive. These queries and its output will be discussed in detail in Results section.



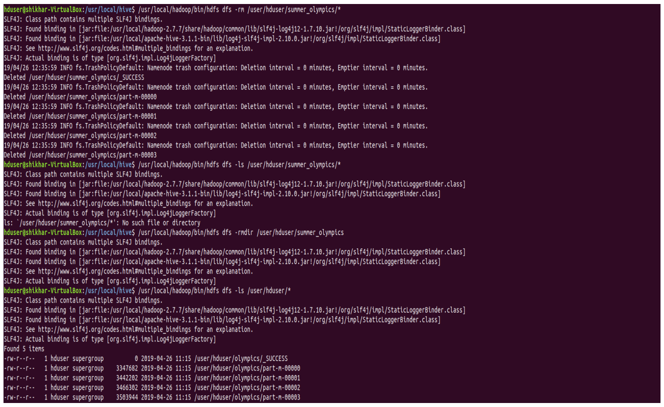
```
mysql> create table summer_olympics(ID int primary key, Name varchar(255), Sex varchar(10), Age double, Height double, Weight double, Team varchar(255), NOC varchar(50), Games varchar(255), Year varchar(10), Sea
son varchar(50), City varchar(50), Sport varchar(255), Event varchar(255), Medal varchar(50));
Query OK, 0 rows affected (0.22 sec)

mysql> show tables;
+----------------------+
| Tables_in_PDA_Project |
+----------------------+
| olympics             |
| summer_olympics      |
+----------------------+
2 rows in set (0.00 sec)

mysql> load data local infile '/home/hduser/OlympicFiltering/output/mapreduceOP' into table summer_olympics fields terminated by ',' lines terminated by '\n' IGNORE 1 LINES;
Query OK, 84307 rows affected (4.67 sec)
Records: 84307  Deleted: 0  Skipped: 0  Warnings: 0

mysql> select * from summer_olympics limit 50;
+----+------------------------------------------+------+------+--------+--------+---------------+------+-------------+------+--------+----------------+--------------+------------------------------------------
-+-------------+
| ID | Name                                     | Sex  | Age  | Height | Weight | Team          | NOC  | Games       | Year | Season | City           | Sport        | Event
 | Medal       |
+----+------------------------------------------+------+------+--------+--------+---------------+------+-------------+------+--------+----------------+--------------+------------------------------------------
-+-------------+
|  1 | A Dijiang                                | M    |   24 |    180 |     80 | China         | CHN  | 1992 Summer | 1992 | Summer | Barcelona      | Basketball   | Basketball Men's Basketball
 | No Medal    |
|  2 | A Lamusi                                 | M    |   23 |    170 |     60 | China         | CHN  | 2012 Summer | 2012 | Summer | London         | Judo         | Judo Men's Extra-Lightweight
 | No Medal    |
| 12 | Jyri Tapani Aalto                        | M    |   31 |    172 |     70 | Finland       | FIN  | 2000 Summer | 2000 | Summer | Sydney         | Badminton    | Badminton Men's Singles
 | No Medal    |
| 13 | Minna Maarit Aalto                       | F    |   30 |    159 |   55.5 | Finland       | FIN  | 1996 Summer | 1996 | Summer | Atlanta        | Sailing      | Sailing Women's Windsurfer
 | No Medal    |
| 17 | Paavo Johannes Aaltonen                  | M    |   28 |    175 |     64 | Finland       | FIN  | 1948 Summer | 1948 | Summer | London         | Gymnastics   | Gymnastics Men's Individual All-Around
 | Bronze      |
| 18 | Timo Antero Aaltonen                     | M    |   31 |    189 |    130 | Finland       | FIN  | 2000 Summer | 2000 | Summer | Sydney         | Athletics    | Athletics Men's Shot Put
 | No Medal    |
| 22 | Andreea Aanei                            | F    |   22 |    170 |    125 | Romania       | ROU  | 2016 Summer | 2016 | Summer | Rio de Janeiro | Weightlifting | Weightlifting Women's Super-Heavyweight
 | No Medal    |
| 23 | Fritz Aanes                              | M    |   22 |    187 |     89 | Norway        | NOR  | 2000 Summer | 2000 | Summer | Sydney         | Wrestling    | Wrestling Men's Light-Heavyweight
|  Greco-Roman |
| 28 | Jan-Erik Aarberg                         | M    |   43 |    170 |     77 | Norway        | NOR  | 1968 Summer | 1968 | Summer | Mexico City    | Sailing      | Sailing Mixed Three Person Keelboat
 | No Medal    |
| 30 | Pepijn Aardewijn                         | M    |   26 |    189 |     72 | Netherlands   | NED  | 1996 Summer | 1996 | Summer | Atlanta        | Rowing       | Rowing Men's Lightweight Double Sculls
 | Silver      |
| 31 | Evald rma (rman-)                        | M    |   24 |    174 |     70 | Estonia       | EST  | 1936 Summer | 1936 | Summer | Berlin         | Athletics    | Athletics Men's Pole Vault
 | No Medal    |
| 33 | Mika Lauri Aarnikka                      | M    |   24 |    187 |     76 | Finland       | FIN  | 1992 Summer | 1992 | Summer | Barcelona      | Sailing      | Sailing Men's Two Person Dinghy
 | No Medal    |
| 34 | Jamale (Djamel-) Aarrass (Ahrass-)       | M    |   30 |    187 |     76 | France        | FRA  | 2012 Summer | 2012 | Summer | London         | Athletics    | Athletics Men's 1
```

Fig. 12. Summer Olympics Table Created in MySQL


Activities
Terminal
Sat 14:04
ie
hduser@shikhar-VirtualBox: /usr/local/hive
File Edit View Search Terminal Help

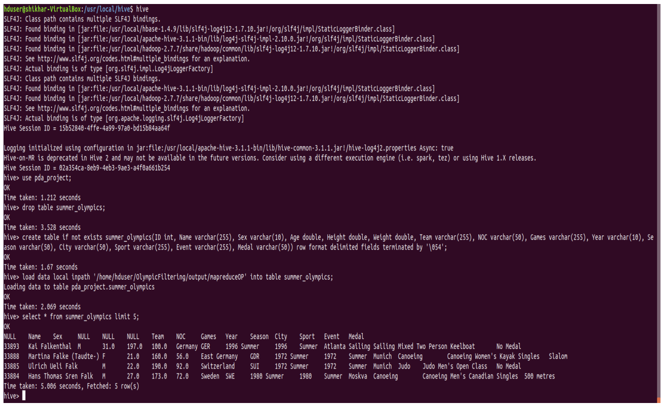
```
hduser@shikhar-VirtualBox:/usr/local/hive$ sqoop import --connect jdbc:mysql://127.0.0.1/PDA_Project --username root --password shikhar96969 --table summer_olympics --hive-import --hive-table summer_olympics --h
ive-database PDA_Project --columns "ID,Name,Sex,Age,Height,Weight,Team,NOC,Games,Year,Season,City,Sport,Event,Medal"
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
19/04/26 12:21:07 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
19/04/26 12:21:07 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
19/04/26 12:21:07 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
19/04/26 12:21:07 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
19/04/26 12:21:08 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/04/26 12:21:08 INFO tool.CodeGenTool: Beginning code generation
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Fri Apr 26 12:21:09 IST 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be establ
ished by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting
 useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
19/04/26 12:21:09 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `summer_olympics` AS t LIMIT 1
19/04/26 12:21:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `summer_olympics` AS t LIMIT 1
19/04/26 12:21:10 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-hduser/compile/bbabc1255414c66f7c8c32a9553c056f/summer_olympics.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/04/26 12:21:17 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hduser/compile/bbabc1255414c66f7c8c32a9553c056f/summer_olympics.jar
19/04/26 12:21:17 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/04/26 12:21:17 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/04/26 12:21:17 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/04/26 12:21:17 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/04/26 12:21:17 INFO mapreduce.ImportJobBase: Beginning import of summer_olympics
19/04/26 12:21:18 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/04/26 12:21:21 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/04/26 12:21:21 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/26 12:21:22 ERROR tool.ImportTool: Import failed: org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://localhost:54310/user/hduser/summer_olympics already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:266)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:139)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1290)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1287)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1762)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1287)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1308)
        at org.apache.sqoop.mapreduce.ImportJobBase.doSubmitJob(ImportJobBase.java:200)
        at org.apache.sqoop.mapreduce.ImportJobBase.runJob(ImportJobBase.java:173)
        at org.apache.sqoop.mapreduce.ImportJobBase.runImport(ImportJobBase.java:270)
        at org.apache.sqoop.manager.SqlManager.importTable(SqlManager.java:692)
        at org.apache.sqoop.manager.MySQLManager.importTable(MySQLManager.java:127)
        at org.apache.sqoop.tool.ImportTool.importTable(ImportTool.java:520)
        at org.apache.sqoop.tool.ImportTool.run(ImportTool.java:628)
        at org.apache.sqoop.Sqoop.run(Sqoop.java:147)
```

Fig. 15. Second Sqoop Trial



```
hduser@shikhar-VirtualBox:/usr/local/hive$ /usr/local/hadoop/bin/hdfs dfs -rm /user/hduser/summer_olympics/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
19/04/26 12:35:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hduser/summer_olympics/_SUCCESS
19/04/26 12:35:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hduser/summer_olympics/part-m-00000
19/04/26 12:35:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hduser/summer_olympics/part-m-00001
19/04/26 12:35:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hduser/summer_olympics/part-m-00002
19/04/26 12:35:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hduser/summer_olympics/part-m-00003
hduser@shikhar-VirtualBox:/usr/local/hive$ /usr/local/hadoop/bin/hdfs dfs -ls /user/hduser/summer_olympics/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
ls: `/user/hduser/summer_olympics/*': No such file or directory
hduser@shikhar-VirtualBox:/usr/local/hive$ /usr/local/hadoop/bin/hdfs dfs -rmdir /user/hduser/summer_olympics
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hduser@shikhar-VirtualBox:/usr/local/hive$ /usr/local/hadoop/bin/hdfs dfs -ls /user/hduser/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 5 items
-rw-r--r--   1 hduser supergroup          0 2019-04-26 11:15 /user/hduser/olympics/_SUCCESS
-rw-r--r--   1 hduser supergroup    3347682 2019-04-26 11:15 /user/hduser/olympics/part-m-00000
-rw-r--r--   1 hduser supergroup    3442202 2019-04-26 11:15 /user/hduser/olympics/part-m-00001
-rw-r--r--   1 hduser supergroup    3466302 2019-04-26 11:15 /user/hduser/olympics/part-m-00002
-rw-r--r--   1 hduser supergroup    3503944 2019-04-26 11:15 /user/hduser/olympics/part-m-00003
```

Fig. 16. Manually Deleting Target Directory



```
hduser@shikhar-VirtualBox:/usr/local/hive$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase-1.4.9/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.1-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 15b52840-4ffe-4a99-97a0-bd15b84aa64f

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.1-bin/lib/hive-common-3.1.1.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = 02a354ca-8eb9-4eb3-9ae3-a4f0a661b254
hive> use pda_project;
OK
Time taken: 1.212 seconds
hive> drop table summer_olympics;
OK
Time taken: 3.528 seconds
hive> create table if not exists summer_olympics(ID int, Name varchar(255), Sex varchar(10), Age double, Height double, Weight double, Team varchar(255), NOC varchar(50), Games varchar(255), Year varchar(10), Se
ason varchar(50), City varchar(50), Sport varchar(255), Event varchar(255), Medal varchar(50)) row format delimited fields terminated by '\054';
OK
Time taken: 1.67 seconds
hive> load data local inpath '/home/hduser/OlympicFiltering/output/mapreduceOP' into table summer_olympics;
Loading data to table pda_project.summer_olympics
OK
Time taken: 2.069 seconds
hive> select * from summer_olympics limit 5;
OK
NULL    Name    Sex     NULL    NULL    NULL    Team    NOC     Games   Year    Season  City    Sport   Event   Medal
33893   Kai Falkenthal  M       31.0    197.0   100.0   Germany GER     1996 Summer     1996    Summer  Atlanta Sailing Sailing Mixed Two Person Keelboat       No Medal
33888   Martina Falke (Taudte-) F       21.0    160.0   56.0    East Germany    GDR     1972 Summer     1972    Summer  Munich  Canoeing        Canoeing Women's Kayak Singles   Slalom
33885   Ulrich Ueli Falk        M       22.0    190.0   92.0    Switzerland     SUI     1972 Summer     1972    Summer  Munich  Judo    Judo Men's Open Class   No Medal
33884   Hans Thomas Sren Falk   M       27.0    173.0   72.0    Sweden  SWE     1980 Summer     1980    Summer  Moskva  Canoeing        Canoeing Men's Canadian Singles  500 metres
Time taken: 5.006 seconds, Fetched: 5 row(s)
hive> 
```

Fig. 17. Inserting Data in Hive Table

Step 11: In this last step, three .pig files were created in which pig latin queries were written and executed on MapReduce output dataset which was pulled in these .pig files. The output from all .pig files was stored in three different folders. All three .pig files and output files are present in the compresed folder uploaded.

IV. Evaluation and Results

In this section, output from all the tasks performed are discussed in detail.

A. R Processing - MySQL Data Loading - MapReduce

After downloading data from Kagggle, a short code was written in R to clean it. The R Code is used to perform pre-processing task is present in the compressed folder. After a little bit of pre-processing, the data is then put in shared folder to use it in Ubuntu VM. In Ubuntu Vm, olympic data file is further moved from the shared directory to hduser and then moved to HDFS using '-copyFromLocal' command. MapReduce task was built in eclipse for filtering Olympic dataset for just summer Olympics. This MapReduce task was built to filter the Season column in dataset for just summer (file present in compressed folder) and this dataset will further be fed to other databases for analyzing. After filtering data was left with 84307 records in 15 columns holding details of summer Olympics from year 1896 to 2016. This data was further organized a little bit as the MapReduce output was having leading spaces and header was also missing.

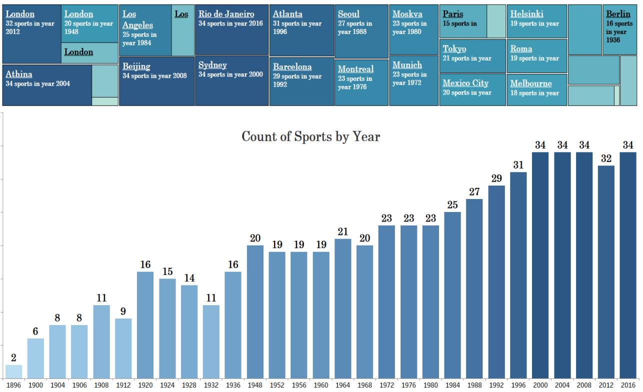


Fig. 18. Count of Sports by Year and City

B. HBase

HBase in this project is only used to store the MapReduce output in it. To store MapReduce output in HBase, data was first moved to HDFS and then pulled in HBase. No other task was performed in HBase.

C. Hive

There were few errors while sqooping data from MySQL to Hive (as explained in Methodology section), due to which sqoop command was not working and a table was manually created in Hive in which data was manually inserted. Four queries were written in Hive to analyze Olympic data.

1) Query 1: First query which was written in Hive gives an output of total number of sports for each summer Olympic held. The output from this query contained fields like year, city and count of sports which was grouped by city and year. Screenshot has already been attached in Methodology section. Out files is present in compressed folder. Fig. 18 represents the output from this query where we can clearly see that the number of sports increased through the year and in the last summer Olympic held in year 2016, events related to 34 sports were organized. The color in charts reflect the number of sports and higher the number of sports, darker is the color. Similar color logic has been followed for top chart as well.

2) Query 2: In second Hive query, count of players who participated in Olympoic games from year 1896 to 2016 are shown. In the Fig. 19 we can clearly get an idea that the

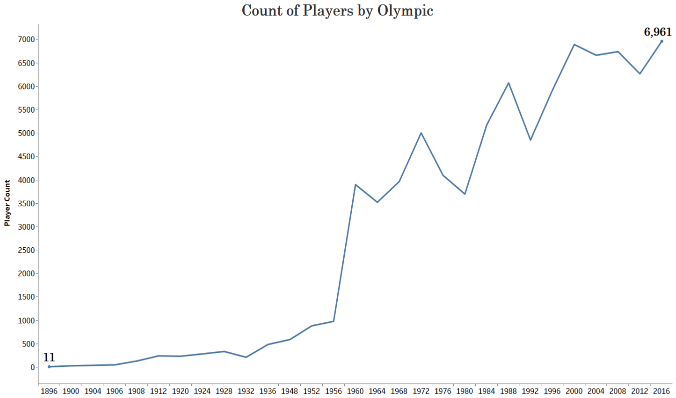


Fig. 19. Count of Players by Year

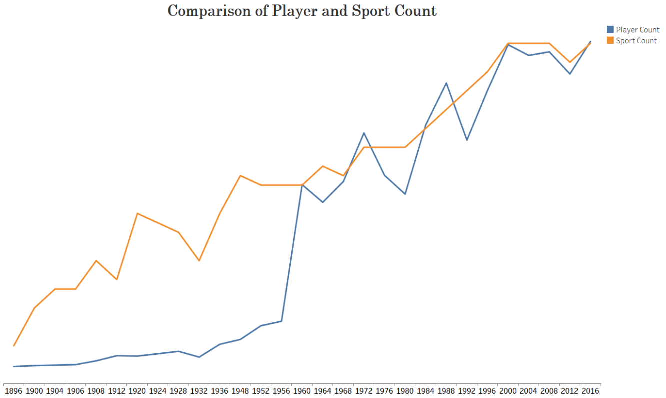


Fig. 20. Sport Count vs. Player Count

number of players participating in summer Olympic games has increased since its inception. Also, we can see that there had been a sudden increase in the number of players participating from year 1956 to 1960. The chart also shows that first Olympic game started with just 11 players.

3) Query 3: The third query shows a comparison of count of players and count of sports. This output was achieved by joining output from first and second query in Hive. In Fig. 20 we can see that there is a steady increase in both numbers through the years till 2016. However, the interesting thing to notice here is that although there is a sudden increase in number of players from year 1956 to 1960, the number of sports organized in these years remains the same.

4) Query 4: Fourth query written in Hive filters data just for last four summer Olympics i.e. summer Olympics held in the year 2016, 2012, 2008 and 2004. A gender wise and country wise analysis was carried out on top of this filtered dataset. Fig. 21 shows an exploratory analysis based on gender and country for last four summer Olympics. Major observations after bulding charts on this dataset are: The number of medals won by females are less than number of medals won by males in each of the four years and it is observed that USA from past four years tops the chart bagging maximum number of medals followed by Russia, China, Germany and Australia.

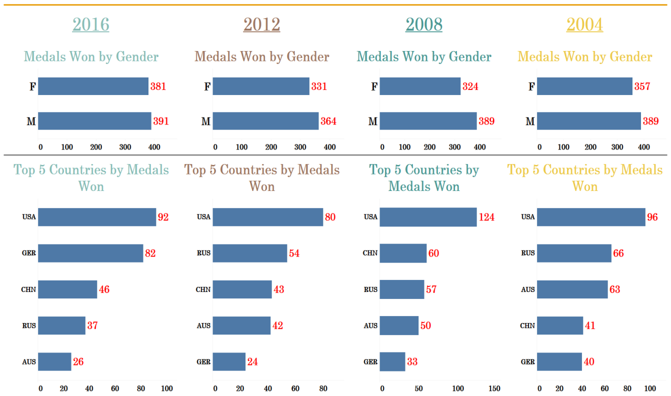


Fig. 21. Last 4 Summer Olympics

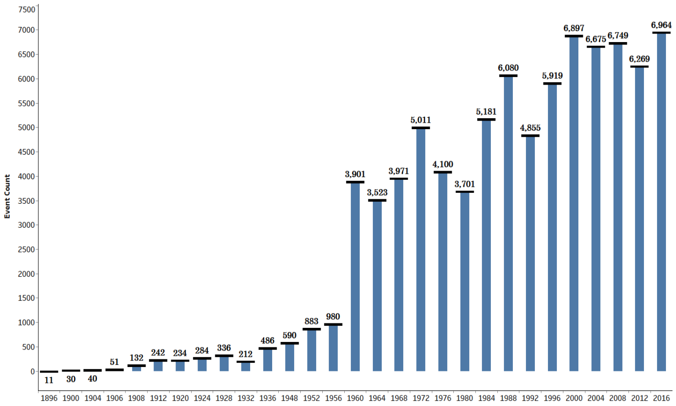


Fig. 22. Number of Events by Olympic

D. Pig

At first Pig was launched in distributed mode data was pulled into Pig from HDFS, but then it was noticed that when while pulling data from HDFS, Pig was trying to connect to server and every time the script was executed it iterated 10 times to try to connect to server which in this case is not possible. So, rather than launching pig in distributed mode, it was launched in local mode and MapReduce output file was then pulled form local system. A total of 3 queries were written in 3 different .pig files and similarly output was stored in 3 different folders.

1) Query 1: First query written in Pig was to get an idea about the number of total events which tool place in each Olympic across multiple sports. Here events for an example refer to different types of races like 100m, 400m etc. under sport running. Like in one of the Hive queries, here also we can see that after year 1956, there had been a considerable rise in number of events with last Olympic in year 2016 witnessing the highest ever number of events. Refer Fig. 22

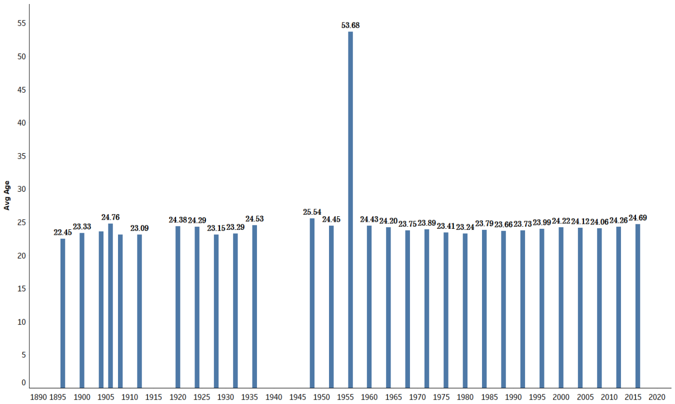


Fig. 23. Avg. Age by Olympic

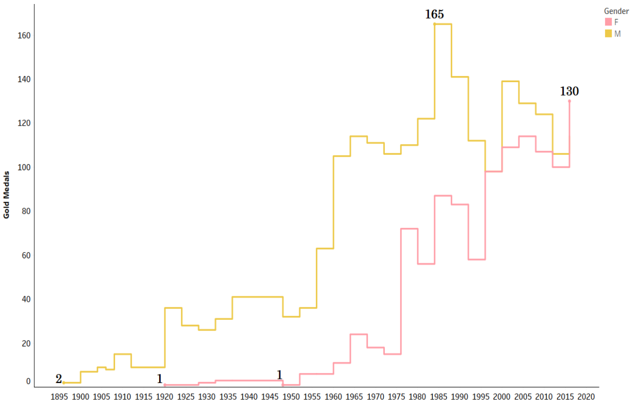


Fig. 24. Avg. Age by Olympic

2) Query 2: Second query checks the average age for each Olympic. Surprisingly, average age for year 1956 Olympic is very high as compared to other Olympics (Fig. 23). For other Olympics, average age is almost the same.

3) Query 3: The third query digs into data to find out number of gold medals by gender. In Fig. 24 we can clearly see two patterns, one showing that there is an increase in medals won by males in females both over the period of different Olympics starting from year 1896 to 2016. Second pattern which can be observed is that the medals won by males in almost every Olympic is more than that of females. In year 1960, there was a considerable increase in number of medals won by makes whereas, year 1976 witnessed the same trend for females.

V. Conclusion and Future Work

As the world is moving towards an era where data will continue to grow exponentially, it has become mandatory to be able to and be prepared to store, manage and process this huge volume of data. Storing data on distributed platforms rather on traditional relational databases is the new approach which many organizations are currently following. This project leverages the power of some popularly known Big Data open source technologies and systematically tries to understand their use whilst getting to know some common errors which could be faced while implementing these technologies for handling data.

Olympic dataset which was downloaded from Kaggle was used for this project which was first pre-processed a little but before using it for data analysis in Ubuntu. This dataset was first loaded into MySQL and then sqooped to HDFS after which a MapReduce filtering job was applied on it. Output from MapReduce was organized and was stored in a directory. This output dataset was further pulled in HBase where it was just stored. Later, MapReduce output data was tried to be pulled in Hive using Sqoop but faced some errors due to which it was then manually pulled in Hive after creating a table, following which four Hive queries were built on this data. Lastly, MapReduce output data was then pulled in Pig and three queries were built on it.

Major observations while building charts on query output from Hive and Pig are: All charts are following an increasing trend which particularly means that every Olympic starting from year 1896 saw an increase in number of participants, number of events, number of sports and number of medals won; for most of the charts year 1960 saw great jump in above mentioned numbers somehow; as far as last four Olympics are concerned, USA tops the chart in bagging most number of medals followed by Russia, China, Germany and Australia. Future work proposed would be to analyse data using HBase, building Pig queries for data present in HDFS rather than using Pig in local mode and using sqoop in all cases for importing data from MySQL into Hive or Pig.

References

[1] 120 years of Olympic history: athletes and results. [Online]. Available: https://kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results. [Accessed: 28-Apr-2019].

[2] N. Spivey and N. J. Spivey, The Ancient Olympics. Oxford University Press, 2005.

[3] A. Guttmann, The Olympics: A History of the Modern Games. University of Illinois Press, 2002.

[4] G. Waitt, Social impacts of the Sydney Olympics, Annals of Tourism Research, vol. 30, no. 1, pp. 194215, Jan. 2003.

[5] A. Thusoo et al., Hive: A Warehousing Solution over a Map-reduce Framework, Proc. VLDB Endow., vol. 2, no. 2, pp. 16261629, Aug. 2009.

[6] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, Pig Latin: A Not-so-foreign Language for Data Processing, in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2008, pp. 10991110.

[7] D. Vohra, Using Apache Sqoop, in Pro Docker, D. Vohra, Ed. Berkeley, CA: Apress, 2016, pp. 151183.

[8] M. S. S. Aravinth and A. H. Begam, An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing, vol. 1, no. 10, p. 4.