IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DATA MINING IN ELITE SPORTS USING APACHE HADOOP AND APACHE PIG

D. Shireesha*, D. Srikanth

Associate Professor Department of Computer Science & Engineering Guru Nanak Institutions Technical Campus
Sr. Sofware Engineer , Accenture

DOI: 10.5281/zenodo.154248

ABSTRACT

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us. Using the information kept in the social network like Face book, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.

In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software's can be written to interact with the database, process the required data and present it to the users for analysis purpose.

KEYWORDS: Big data , Heavy server, Hadoop, Unstructured.

INTRODUCTION

The term big data refers to the data that is generating around us everyday life. It is generally exceeds the capacity of normal conventional traditional databases. Big data is recycled ubiquitously at the present in disseminated archetype on web. It is the group of collections of massive volume of data. That's the reason why big data came into picture in the real time business analysis of processing data.

Multimedia and individuals with smart phones and on social network sites will continue to fuel exponential growth. Big data is large pools of data that can be captured, communicated; aggregated, stored, and analyzed is now part of every sector and function of the global economy. Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn't take place without data.

RELATED WORKS

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets[1]. Hadoop Pig, Hive, and many other projects provide the foundation for storing and processing large amounts of data in an efficient way. Most of the time, it is not possible to perform all required processing with a single MapReduce, Pig, or Hive job. Multiple MapReduce, Pig, or Hive jobs often need to be chained together, producing and consuming intermediate data and coordinating their flow of execution[3].

As developers started doing more complex processing using Hadoop, multistage Hadoop jobs became common. This lead to several ad hoc solutions to manage the execution and interdependency of these multiple Hadoop jobs. Some developers wrote simple shell scripts to start one Hadoop job after the other. Others used Hadoop's Job Control class, which executes multiple MapReduce jobs using topological sorting[4].

ISSN: 2277-9655
Impact Factor: 4.116

PROPOSED SYSTEM

It was clear that there was a need for a general-purpose system to run multistage
Hadoop jobs with the following requirements

- It should use an adequate and well-understood programming model to facilitate its adoption and to reduce developer ramp-up time.
- It should be easy to troubleshoot and recover jobs when something goes wrong.
- It should be extensible to support new types of jobs.
- It should scale to support several thousand concurrent jobs.
- Jobs should run in a server to increase reliability.

Proposed System Advantages

- Lazy evaluation: unless you do not produce an output file or does not output any message, it does not get evaluated. This has an advantage in the logical plan, it could optimize the program beginning to end and optimizer could produce an efficient plan to execute.

DESIGN ENGINEERING

Activity Diagram

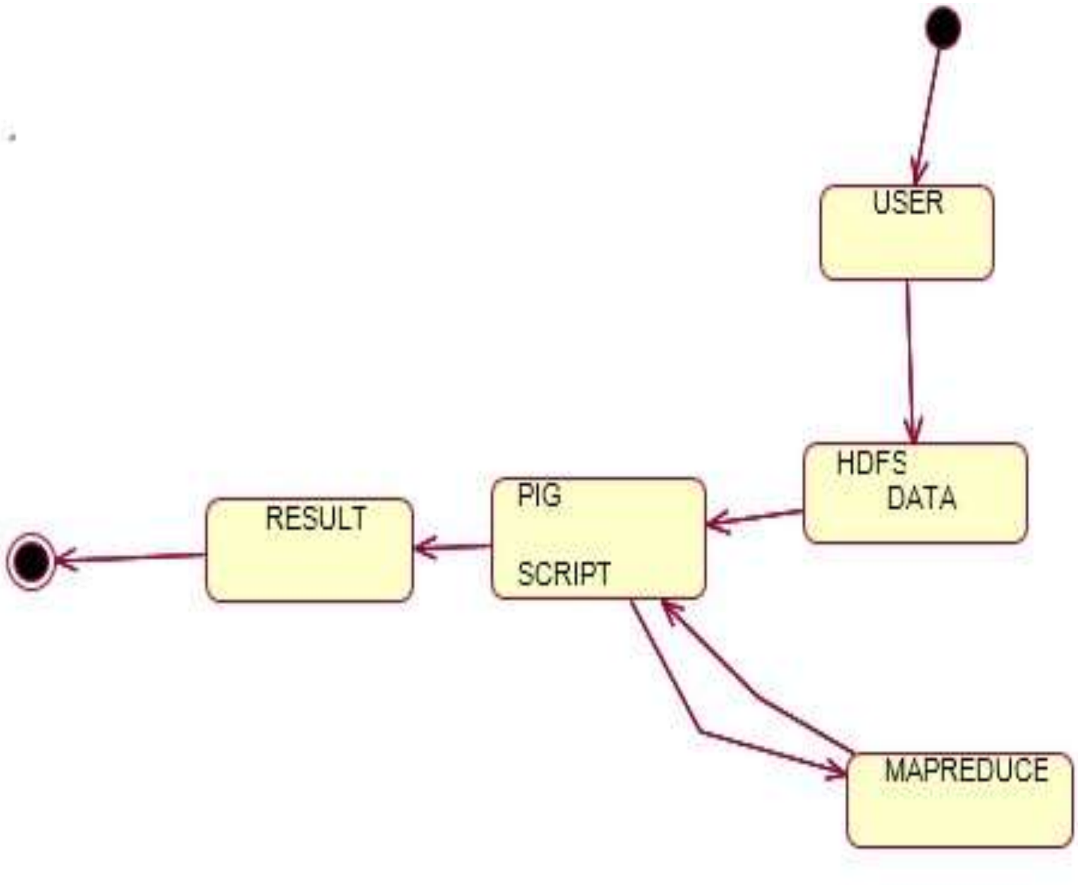


Figure 4.1 ACTIVITY DIAGRAM

Use Case Diagram

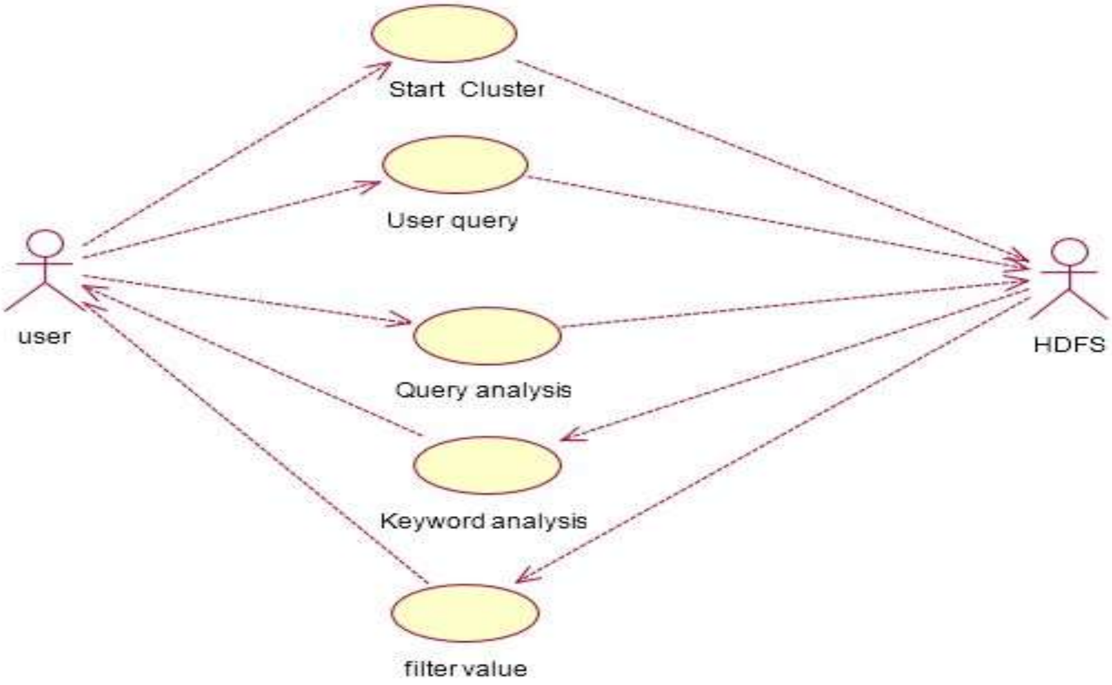


Figure 4.2 USE CASE DIAGRAM

Sequence Diagram

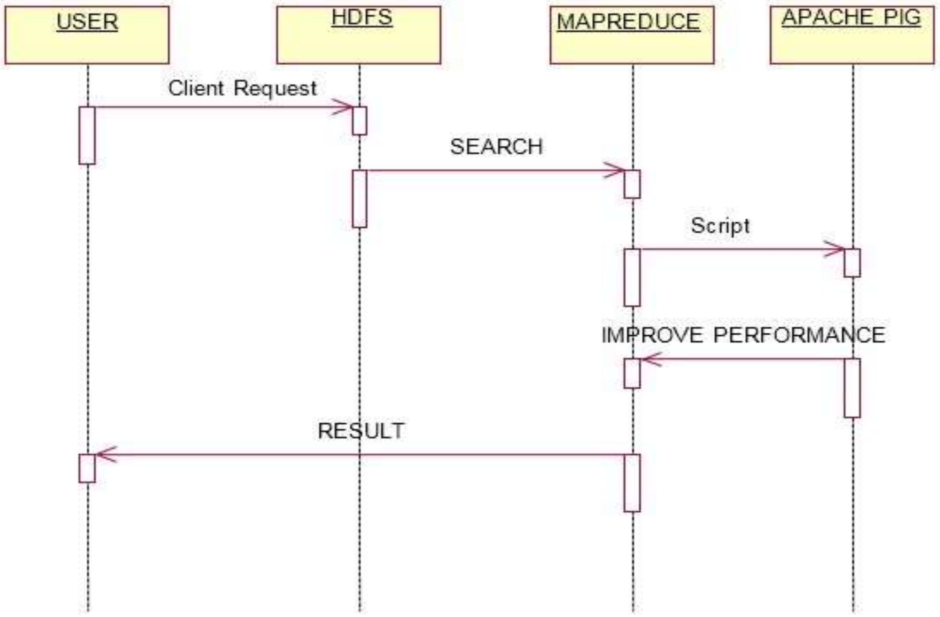


Figure 4.3 SEEQUENCE DIAGRAM

SYSTEM ARCHITECTURE

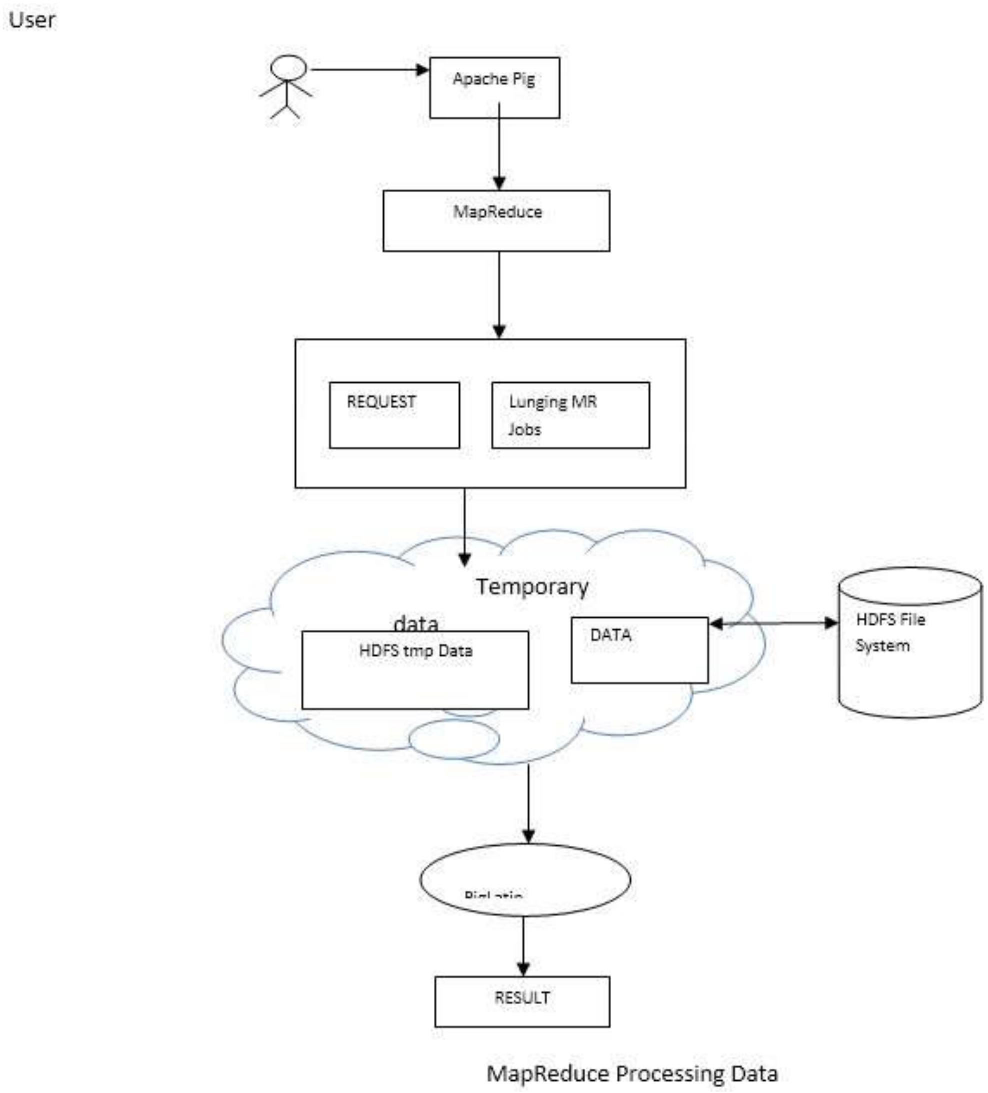
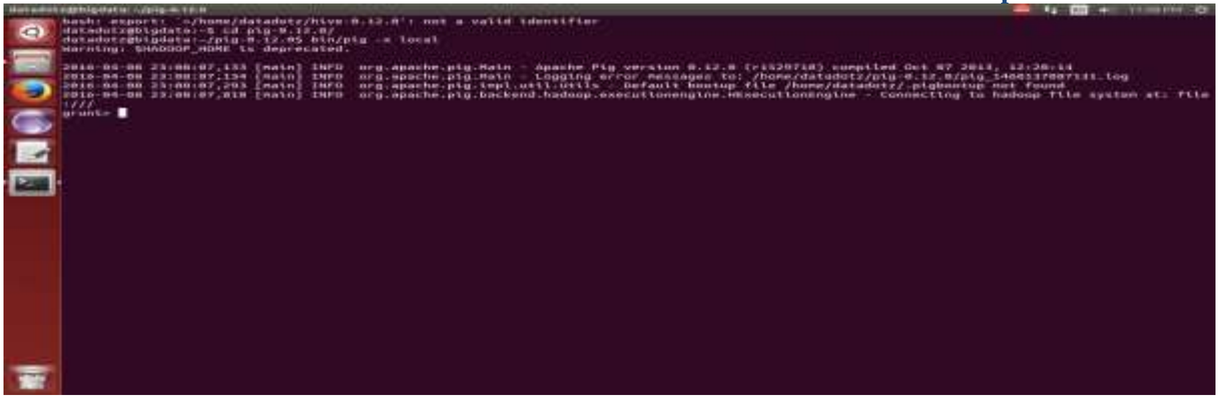


Figure 4.4 SYSTEM ARCHITECTURE DIAGRAM

The systems architect establishes the basic structure of the system, defining the essential core design features and elements that provide the framework for all that follows, and are the hardest to change later. The systems architect provides the architects view of the users' vision for what the system needs to be and do, and the paths along which it must be able to evolve, and strives to maintain the integrity of that vision as it evolves during detailed design and implementation.

SCREENSHOTS

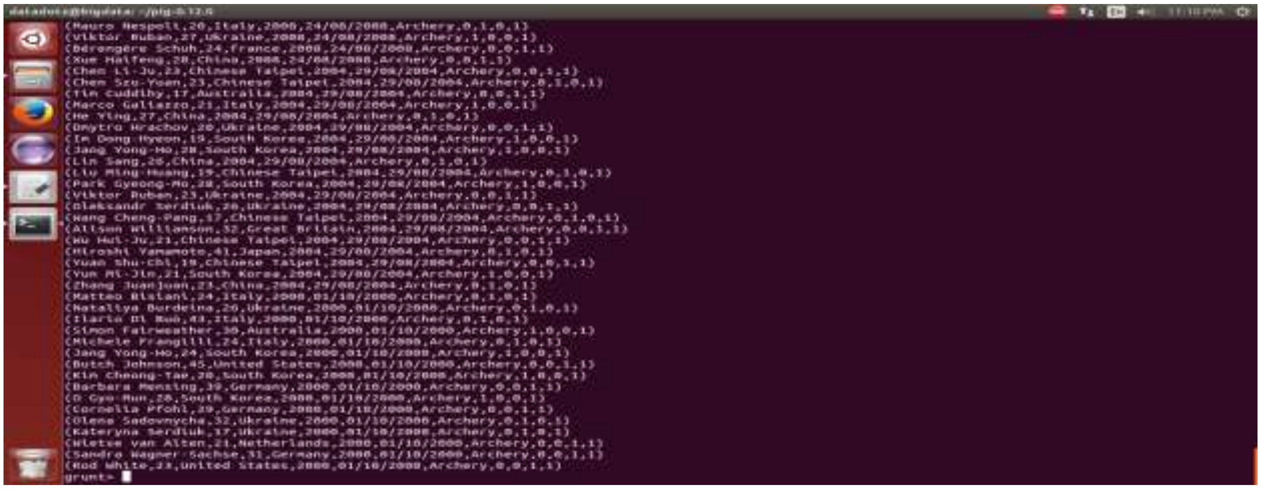
Entering Into Pig Local Machine



```
bash: export: `=/home/datadotz/hive-0.12.0': not a valid identifier
datadotz@bigdata:~$ cd pig-0.12.0/
datadotz@bigdata:~/pig-0.12.0$ bin/pig -x local
Warning: $HADOOP_HOME is deprecated.

2016-04-08 23:08:07,133 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0 (r1529718) compiled Oct 07 2013, 12:20:14
2016-04-08 23:08:07,134 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/datadotz/pig-0.12.0/pig_1460137087131.log
2016-04-08 23:08:07,293 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/datadotz/.pigbootup not found
2016-04-08 23:08:07,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file
:///
grunt> 
```

Figure 5.1 : ENTERING INTO PIG LOCAL MACHINE

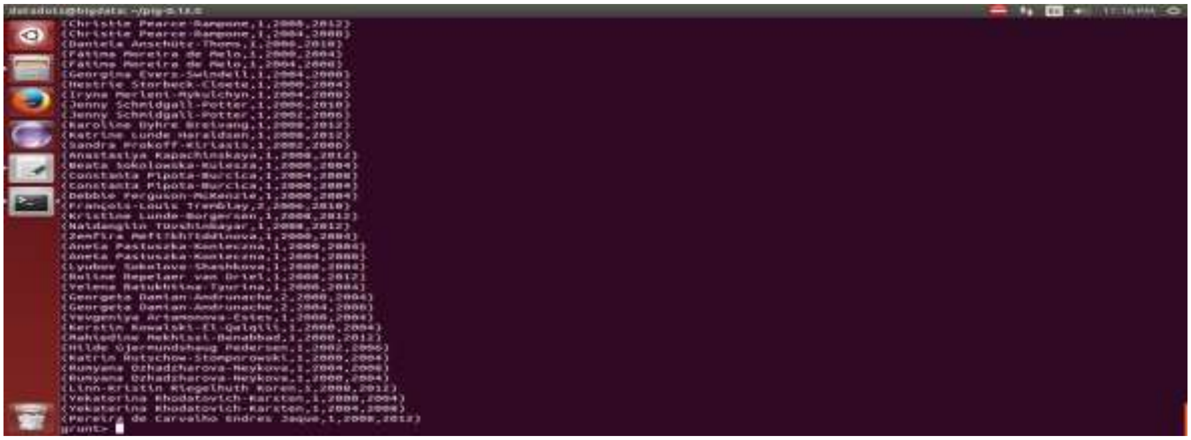
RESULTANTTABLE



```
(Mauro Nespoli,20,Italy,2008,24/08/2008,Archery,0,1,0,1)
(Viktor Ruban,27,Ukraine,2008,24/08/2008,Archery,1,0,0,1)
(Bérengère Schuh,24,France,2008,24/08/2008,Archery,0,0,1,1)
(Xue Haifeng,28,China,2008,24/08/2008,Archery,0,0,1,1)
(Chen Li-Ju,23,Chinese Taipei,2004,29/08/2004,Archery,0,0,1,1)
(Chen Szu-Yuan,23,Chinese Taipei,2004,29/08/2004,Archery,0,1,0,1)
(Tim Cuddihy,17,Australia,2004,29/08/2004,Archery,0,0,1,1)
(Marco Galiazzo,21,Italy,2004,29/08/2004,Archery,1,0,0,1)
(He Ying,27,China,2004,29/08/2004,Archery,0,1,0,1)
(Dmytro Hrachov,20,Ukraine,2004,29/08/2004,Archery,0,0,1,1)
(Im Dong-Hyeon,19,South Korea,2004,29/08/2004,Archery,1,0,0,1)
(Jang Yong-Ho,28,South Korea,2004,29/08/2004,Archery,1,0,0,1)
(Lin Sang,26,China,2004,29/08/2004,Archery,0,1,0,1)
(Liu Ming-Huang,19,Chinese Taipei,2004,29/08/2004,Archery,0,1,0,1)
(Park Gyeong-Mo,28,South Korea,2004,29/08/2004,Archery,1,0,0,1)
(Viktor Ruban,23,Ukraine,2004,29/08/2004,Archery,0,0,1,1)
(Oleksandr Serdiuk,28,Ukraine,2004,29/08/2004,Archery,0,0,1,1)
(Wang Cheng-Pang,17,Chinese Taipei,2004,29/08/2004,Archery,0,1,0,1)
(Alison Williamson,32,Great Britain,2004,29/08/2004,Archery,0,0,1,1)
(Wu Hui-Ju,21,Chinese Taipei,2004,29/08/2004,Archery,0,0,1,1)
(Hiroshi Yamamoto,41,Japan,2004,29/08/2004,Archery,0,1,0,1)
(Yuan Shu-Chi,19,Chinese Taipei,2004,29/08/2004,Archery,0,0,1,1)
(Yun Mi-Jin,21,South Korea,2004,29/08/2004,Archery,1,0,0,1)
(Zhang Juanjuan,23,China,2004,29/08/2004,Archery,0,1,0,1)
(Matteo Bisiani,24,Italy,2000,01/10/2000,Archery,0,1,0,1)
(Nataliya Burdeina,26,Ukraine,2000,01/10/2000,Archery,0,1,0,1)
(Ilario Di Buò,43,Italy,2000,01/10/2000,Archery,0,1,0,1)
(Simon Fairweather,30,Australia,2000,01/10/2000,Archery,1,0,0,1)
(Michele Frangilli,24,Italy,2000,01/10/2000,Archery,0,1,0,1)
(Jang Yong-Ho,24,South Korea,2000,01/10/2000,Archery,1,0,0,1)
(Butch Johnson,45,United States,2000,01/10/2000,Archery,0,0,1,1)
(Kim Cheong-Tae,20,South Korea,2000,01/10/2000,Archery,1,0,0,1)
(Barbara Mensing,39,Germany,2000,01/10/2000,Archery,0,0,1,1)
(O Gyo-Mun,28,South Korea,2000,01/10/2000,Archery,1,0,0,1)
(Cornelia Pfohl,29,Germany,2000,01/10/2000,Archery,0,0,1,1)
(Olena Sadovnycha,32,Ukraine,2000,01/10/2000,Archery,0,1,0,1)
(Kateryna Serdiuk,17,Ukraine,2000,01/10/2000,Archery,0,1,0,1)
(Wietse van Alten,21,Netherlands,2000,01/10/2000,Archery,0,0,1,1)
(Sandra Wagner-Sachse,31,Germany,2000,01/10/2000,Archery,0,0,1,1)
(Rod White,23,United States,2000,01/10/2000,Archery,0,0,1,1)
grunt> 
```

Figure 5.2: RESULTANT TABLE

OUTPUT OF FOREACH FILTERED

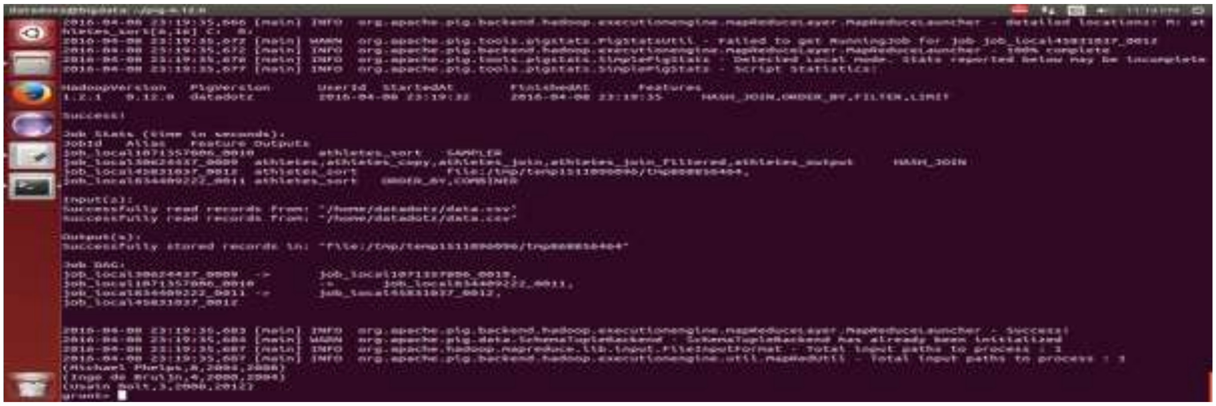


```
(Christie Pearce-Rampone,1,2008,2012)
(Christie Pearce-Rampone,1,2004,2008)
(Daniela Anschütz-Thoms,1,2006,2010)
(Fatima Moreira de Melo,1,2000,2004)
(Fatima Moreira de Melo,1,2004,2008)
(Georgina Evers-Swindell,1,2004,2008)
(Hestrie Storbeck-Cloete,1,2000,2004)
(Iryna Merleni-Mykulchyn,1,2004,2008)
(Jenny Schmidgall-Potter,1,2006,2010)
(Jenny Schmidgall-Potter,1,2002,2006)
(Karoline Dyhre Breivang,1,2008,2012)
(Katrine Lunde Haraldsen,1,2008,2012)
(Sandra Prokoff-Kiriasis,1,2002,2006)
(Anastasiya Kapachinskaya,1,2008,2012)
(Beata Sokolowska-Kulesza,1,2000,2004)
(Constanta Pipota-Burcica,1,2004,2008)
(Constanta Pipota-Burcica,1,2000,2004)
(Debbie Ferguson-McKenzie,1,2000,2004)
(François-Louis Tremblay,2,2006,2010)
(Kristine Lunde-Borgersen,1,2008,2012)
(Naidangiin Tüvshinbayar,1,2008,2012)
(Zemfira Meftakhetdinova,1,2000,2004)
(Aneta Pastuszka-Konieczna,1,2000,2004)
(Aneta Pastuszka-Konieczna,1,2004,2008)
(Lyubov Sokolova-Shashkova,1,2000,2004)
(Roline Repelaer van Driel,1,2008,2012)
(Yelena Ratukhtina-Tyurina,1,2000,2004)
(Georgeta Damian-Andrunache,2,2000,2004)
(Georgeta Damian-Andrunache,2,2004,2008)
(Yevgeniya Artamonova-Estes,1,2000,2004)
(Kerstin Kowalski-El Qalqili,1,2000,2004)
(Mahiedine Mekhissi-Benabbad,1,2008,2012)
(Hilde Gjermundshaug Pedersen,1,2002,2006)
(Katrin Rutschow-Stomporowski,1,2000,2004)
(Rumyana Dzhadzharova-Neykova,1,2004,2008)
(Rumyana Dzhadzharova-Neykova,1,2000,2004)
(Linn-Kristin Riegelhuth Koren,1,2008,2012)
(Yekaterina Khodatovich-Karsten,1,2000,2004)
(Yekaterina Khodatovich-Karsten,1,2004,2008)
(Pereira de Carvalho Endres Jaqué,1,2008,2012)
grunt> 
```

Figure 5.4 OUTPUT OF FOREACH FILTERED

Resultant athletes data is displayed with their gold medals won by the athletes in respective years.

FINAL OUTPUT WITH 3 MEDALS

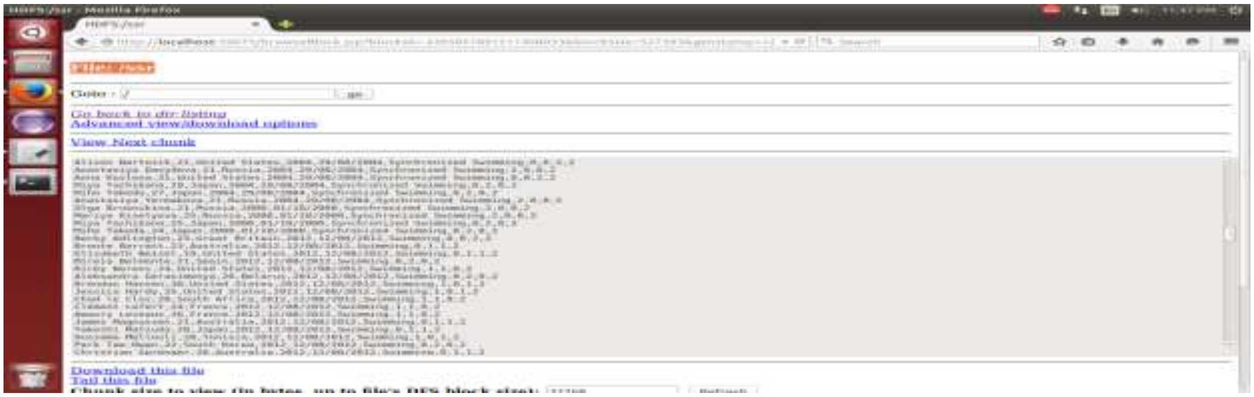


```
datadotz@bigdata: ~/pig-0.12.0
2016-04-08 23:19:35,666 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: at
hletes_sort[4,16] C:  R:
2016-04-08 23:19:35,672 [main] WARN  org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local45831837_0012
2016-04-08 23:19:35,672 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-04-08 23:19:35,676 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2016-04-08 23:19:35,677 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt               FinishedAt              Features
1.2.1   0.12.0  datadotz        2016-04-08 23:19:32     2016-04-08 23:19:35     HASH_JOIN,ORDER_BY,FILTER,LIMIT

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local1071357086_0010        athletes_sort   SAMPLER
job_local386244437_0009 athletes,athletes_copy,athletes_join,athletes_join_filtered,athletes_output     HASH_JOIN
job_local45831837_0012  athletes_sort           file:/tmp/temp1511890096/tmp808856464,
job_local1834480222_0011 athletes_sort  ORDER_BY,COMBINER

Input(s):
Successfully read records from: "/home/datadotz/data.csv"
Successfully read records from: "/home/datadotz/data.csv"

Output(s):
Successfully stored records in: "file:/tmp/temp1511890096/tmp808856464"

Job DAG:
job_local386244437_0009 ->      job_local1071357086_0010,
job_local1071357086_0010        ->      job_local1834480222_0011,
job_local1834480222_0011 ->     job_local45831837_0012,
job_local45831837_0012


2016-04-08 23:19:35,683 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2016-04-08 23:19:35,684 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2016-04-08 23:19:35,687 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2016-04-08 23:19:35,687 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Michael Phelps,8,2004,2008)
(Ingo de Bruijn,4,2000,2004)
(Usain Bolt,3,2008,2012)
grunt> 
```

Figure 5.5 FINAL OUTPUT WITH 3 MEDALS

Resultant sorted data for more than 3 gold medals won by the athletes.

RESULTANT DATA STORED IN HDFS



File: /ssr

Goto : / go

Go back to dir listing
Advanced view/download options

View Next chunk

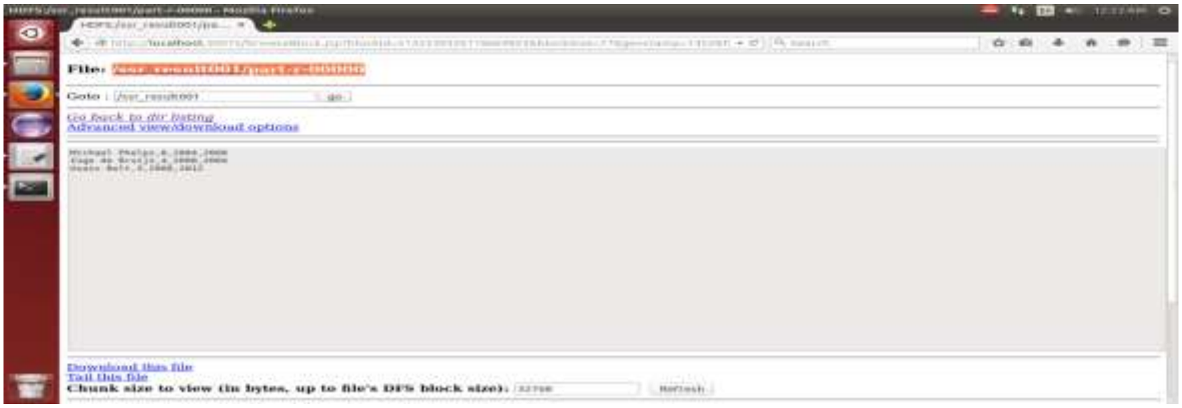
```
Alison Bartosik,21,United States,2004,29/08/2004,Synchronized Swimming,0,0,2,2
Anastasiya Davydova,21,Russia,2004,29/08/2004,Synchronized Swimming,2,0,0,2
Anna Kozlova,31,United States,2004,29/08/2004,Synchronized Swimming,0,0,2,2
Miya Tachibana,29,Japan,2004,29/08/2004,Synchronized Swimming,0,2,0,2
Miho Takeda,27,Japan,2004,29/08/2004,Synchronized Swimming,0,2,0,2
Anastasiya Yermakova,21,Russia,2004,29/08/2004,Synchronized Swimming,2,0,0,2
Olga Brusnikina,21,Russia,2000,01/10/2000,Synchronized Swimming,2,0,0,2
Mariya Kiselyova,25,Russia,2000,01/10/2000,Synchronized Swimming,2,0,0,2
Miya Tachibana,25,Japan,2000,01/10/2000,Synchronized Swimming,0,2,0,2
Miho Takeda,24,Japan,2000,01/10/2000,Synchronized Swimming,0,2,0,2
Becky Adlington,23,Great Britain,2012,12/08/2012,Swimming,0,0,2,2
Bronte Barratt,23,Australia,2012,12/08/2012,Swimming,0,1,1,2
Elizabeth Beisel,19,United States,2012,12/08/2012,Swimming,0,1,1,2
Mireia Belmonte,21,Spain,2012,12/08/2012,Swimming,0,2,0,2
Ricky Berens,24,United States,2012,12/08/2012,Swimming,1,1,0,2
Aleksandra Gerasimenya,26,Belarus,2012,12/08/2012,Swimming,0,2,0,2
Brendan Hansen,30,United States,2012,12/08/2012,Swimming,1,0,1,2
Jessica Hardy,25,United States,2012,12/08/2012,Swimming,1,0,1,2
Chad le Clos,20,South Africa,2012,12/08/2012,Swimming,1,1,0,2
Clement Lefert,24,France,2012,12/08/2012,Swimming,1,1,0,2
Amaury Leveaux,26,France,2012,12/08/2012,Swimming,1,1,0,2
James Magnussen,21,Australia,2012,12/08/2012,Swimming,0,1,1,2
Takeshi Matsuda,28,Japan,2012,12/08/2012,Swimming,0,1,1,2
Oussama Mellouli,28,Tunisia,2012,12/08/2012,Swimming,1,0,1,2
Park Tae-Hwan,22,South Korea,2012,12/08/2012,Swimming,0,2,0,2
Christian Sprenger,26,Australia,2012,12/08/2012,Swimming,0,1,1,2
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

Figure 5.6 RESULTANT DATA STORED IN HDFS

Open browser, where we can see the resultant athletes data set in hdfs**.**

RESULTANT TOP ATHLETES DATA IN HDFS



File: /ssr_result001/part-r-00000

Goto : /ssr_result001 go

Go back to dir listing
Advanced view/download options

```
Michael Phelps,8,2004,2008
Ingo de Bruijn,4,2000,2004
Usain Bolt,3,2008,2012
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

Figure 5.7 RESULTANT TOP ATHLETES DATA IN HDFS

We have created the folder as ssr_result001 in which we see the top athletes data set who one more than 3 gold medals in Olympics.

CONCLUSION

While this has focused on one set of data in one Hadoop cluster, you can configure the Hadoop cluster to be one unit of a larger system, connecting it to other systems of Hadoop clusters or even connecting it to the cloud for broader interactions and greater data access. With this big data infrastructure, and its data analytics application, the learning rate of the implemented algorithms will increase with more access to data at rest or to streaming data, which will allow you to make better, quicker, and more accurate business decisions.

FUTURE ENHANCEMENT

With the rise of Apache Hadoop, a next-generation enterprise data architecture is emerging that connects the systems powering business transactions and business intelligence. Hadoop is uniquely capable of storing, aggregating, and refining multi-structured data sources into formats that fuel new business insights. Apache Hadoop is fast becoming the defector platform for processing Big Data.

REFERENCES

[1] Schmidt, Lars-Henrik (1996). "Commonness across Cultures". In Balslev, Anindita Niyogi. Cross-cultural Conversation: Initiation. Oxford University Press. pp. 119–32. ISBN 978-0-7885-0308-5.
[2] Prieto Gutiérrez, J. J. (2011). Herramientas para el análisis y monitoreo en Redes Sociales. International Review of Information Ethics, 16, 12 [1]
[3] Lamont, Judith (2011-07-05). "Text analytics finds dynamic growth in e-discovery and customer feedback". KMWorld. Retrieved 2012-04-18.
[4] Kite, Shane (2011-06-01). "Social CRM's a Tough, Worthy Goal". Bank Technology News. Retrieved 2012-04-18.
[5] Thompson, Bob (2009-05-17). "Can You Hear Me Now? Top Five Voice of Customer Pitfalls". CustomerThink. Retrieved 2012-04-18.
[6] Mood of the Nation (detecting Mood and Affect on Twitter). geopatterns.enm.bris.ac.uk