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].
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

Sequence Diagram

Figure 4.2 USE CASE DIAGRAM

Figure 4.3 SEEQUENCE DIAGRAM
SYSTEM ARCHITECTURE

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
Resultant athletes data is displayed with their gold medals won by the athletes in respective years.
FINAL OUTPUT WITH 3 MEDALS

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

RESULTANT DATA STORED IN HDFS

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

RESULTANT TOP ATHLETES DATA IN HDFS
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

[6] Mood of the Nation (detecting Mood and Affect on Twitter). geopatterns.enm.bris.ac.uk