Olympics Big Data Prognostications

Arushi Jain, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India
Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India

ABSTRACT

Data is continuously snowballing over the years, gradually a huge growth is seen in data to store and tame to yield meticulous result. It gives rise to a concept nowadays, reckoned as big data analytics. With the summer Olympics at Rio de Janeiro, Brazil in the year 2016 round the corner, we, the authors have implemented a mathematical model by implementing efficient map reduce program to predict the number of medals each country might bag at the games. Based on a number of factors such as historical performance of the country in terms of medals won, the performance of athletes, financial scenario in the country, fitness levels and nutrition of athletes along with familiarity to the playing conditions can be used to come up with a reliable estimate.

KEYWORDS

Big Data, Flume, Hadoop Distributed File System, Hive, MapReduce, Pig, Sqoop, Unstructured Data

INTRODUCTION

Big data as already become a prominent part of a $64 billion database and analytics souk. Gartner has defined Big Data by giving three characteristics (volume, velocity, variety) prevalently known as 3 V’s of Big Data. Later on, IBM defined one more characteristics (veracity) and provides the theory of 4 V’s. It also includes the 3 V’s given by Gartner. Now-a-days organizations are struggling with fifth V of big data, finding value contained in big data. Analytic souk is leaving no stone unturned to tame the available data that yields interesting results and better insights of industry or organization.

• **Volume**: It refers to the amount of data. When data size surpasses from terabytes to exabytes, it refers to big data. Big firms like Facebook, twitter generate billions of data daily and uses big data analytics to retrieve valuable information when require.

• **Velocity**: The speed at which data is received and processed. It corresponds to data in motion. Real-time applications and Internet-of-things (IOT) requires abrupt response. So in such cases, time for harnessing the data should be in seconds to milliseconds.

• **Variety**: This feature depicts different formats of data. Big data can handle any form of data whether structured, semi-structured or unstructured. Unstructured data can contain same requirements information to structured data but in an unorganized manner not in row-column format.

• **Veracity**: This refers to data in doubt. Veracity in data analytics is a non-desired feature in data. It is the uncertainties and noise present in the data.

• **Value**: Characteristic refers to the intrinsic value contained in big data.

Some of the challenges of big data are:
1. The biggest challenge in big data is to aggregate data from heterogeneous sources and analyzes it to get useful information out of it to improve various aspects of functioning and business process of organizations. The data may come from various social networks, with each having a different format.

2. One of the main characteristics of big data is Autonomous where data source works independently without being dependent on centralized control. For example, World Wide Web generates function correctly without involving other servers.

3. Another big data challenge is complexity as data is collected from the heterogeneous sources.

4. Big data is always evolving; thus evolution of complex data poses a big challenge.

For analyzing large and complex data sets, an application framework is used, known as Hadoop. Hadoop Distributed File System (HDFS) makes use of Distributed File System (DFS) for processing the data. It divides the file into blocks, which is then assigned to the different nodes in the cluster of hadoop framework where the input data is processed with the help of MapReduce programming and the result is written again into HDFS.

Hadoop is not a tool or software; it is a combination of different technologies embedded into a framework that operates simultaneously in Hadoop Architecture. It can operate on single workstations or in cluster of nodes. There are different tools available in market for data analytics. Hadoop is a very popular and open-source processing framework for Big Data. It was an initiative of Yahoo but later it was endorsed by Apache Software Foundation as Apache project. With perpetual success in handling avalanche data, Hadoop has become de facto benchmark for storing, taming, harnessing terabytes, petabytes of data. It facilitates distributed multiple processing of complex and large datasets over cluster of commodity hardware. MapReduce performs data processing. MapReduce is a programming paradigm which can run parallel processing tasks on nodes in a cluster simultaneously and is able to both process and generate really large data sets with ease. It takes input and gives the output in form of key-value pairs. MapReduce is able to achieve all this by simply dividing each problem into two broad phases, the map phase and the reduce phase. The Map phase processes each record sequentially and independently on each node and generates intermediate key-value pairs. The Reduce phase takes the output of the Map phase. It processes and merges all the intermediate values to give the final output, again in form of key-value pairs. The output gets sorted after each phase, thus providing the user with the aggregated output from all nodes in an orderly fashion. This can be easily used for clustering large data sets. MapReduce programs have to be written for the query and the classification takes place on a cluster of interconnected computers, whose numbers may run into thousands. It is a programming framework to process data sets on distributed system. Applications are written in Java Language once and processing of data will take place on different nodes in a cluster. It consists of two phases: Map and Reduce Phase.

The Mapper processes each record sequentially and independently in parallel and generates intermediate key-value pairs.

- Map(k1, v1) → list(k2, v2)

The Reduce phase takes the output of the Map phase. It processes and merges all the intermediate values to give the final output, again in form of key-value pairs.

- Reduce(k2, list(v2)) → list(k3, v3)
Related Content

Movie Analytics for Effective Recommendation System using Pig with Hadoop
Arushi Jain and Vishal Bhatnagar (2016). *International Journal of Rough Sets and Data Analysis* (pp. 82-100).
[www.igi-global.com/article/movie-analytics-for-effective-recommendation-system-using-pig-with-hadoop/150466?camid=4v1a](www.igi-global.com/article/movie-analytics-for-effective-recommendation-system-using-pig-with-hadoop/150466?camid=4v1a)

Development of Image Engineering in the Last 20 Years
[www.igi-global.com/chapter/development-of-image-engineering-in-the-last-20-years/183845?camid=4v1a](www.igi-global.com/chapter/development-of-image-engineering-in-the-last-20-years/183845?camid=4v1a)

Australian Users’ Interactions with E-Services in a Virtual Environment
[www.igi-global.com/chapter/australian-users-interactions-services-virtual/65318?camid=4v1a](www.igi-global.com/chapter/australian-users-interactions-services-virtual/65318?camid=4v1a)
Reversible Data Hiding Scheme for ECG Signal
www.igi-global.com/article/reversible-data-hiding-scheme-for-ecg-signal/206876?camid=4v1a