Data Science: Bigtable, MapReduce and Google File System

Karan B. Maniar[1], Chintan B. Khatri[2]

[1,2]Atharva College of Engineering, University of Mumbai, Mumbai, India

Abstract

Data science is the extension of research findings and drawing conclusions from data[1]. BigTable is built on a few of Google technologies[2]. MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster[3]. Google File System is designed to provide efficient, reliable access to data using large clusters of commodity hardware[4]. This paper will discuss Bigtable, MapReduce and Google File System, along with discussing the top 10 algorithms in data mining in brief.

***Keywords*— Data Science, Bigtable, MapReduce, Google File System, Top 10 algorithms in data mining.**

1. INTRODUCTION

A distributed storage system for managing structured data that is planned to scale to an immense size is called Bigtable. Internet's development has resulted in the establishment and wide usage of many internet based applications[5]. Google started the development of Bigtable in 2003 to address the demands of those applications. The applications that require a large storage volume use Bigtable[6].

For processing and generating large data sets, the programming model that is used is MapReduce. Map and Reduce functions that are ubiquitous in functional programming are the inspiration for MapReduce. It is easy to use. MapReduce programs may not be fast all the time[7].

Google File System (GFS) is a scalable distributed file system for data-intensive applications that are large. There two primary advantages of GFS are the hardware on which it runs is low-priced and it delivers high performance. GFS is a vital tool that empowers to sustain to innovate and strike problems on the Internet[8].

2. BIGTABLE

A Bigtable is a sparse, distributed, persistent multidimensional sorted map. For most databases that are commercial, the scale is too large. Since the scale is large, the cost would be very high. It is a simple data model and supports dynamic control over data layout and format. Bigtable is scalable and self-managing at the same time. Wide applicability, scalability, high performance, and high availability are some of the goals that Bigtable has achieved. There are numerous Google products that are using Bigtable for eg. Google Analytics, Google Finance, Orkut, Personalized Search, Writely, and Google Earth. Bigtable maps two random string values that are row key and column key. Timestamp is mapped into an associated arbitrary byte array, which in turn makes it three dimensional[6]. The API of the Bigtable provides many functions like single row transactions are supported, allowing cells to be used as integer counters and deleting and creating tables and column families. A Bigtable is not a relational database. One of the dimensions of each table is a field for time, permitting for versioning and garbage collection. Bottlenecks and inefficiencies can be eliminated as they emerge[5].

3. MAPREDUCE

The Hadoop programs perform two distinguishable tasks and the term MapReduce refers to them. Map job and reduce job are the two tasks. Breaking down individual elements into tuples after taking a set of data and converting it into another set of data is known as the map job. The output from the map job is then taken as input and those data tuples are converted into a smaller set of tuples. As the name suggests, the map job always precedes the reduce job[9]. There are lots of improvements from Terasort records that are: Shuffle 30%+, Merge improvements. MapReduce also enables to reuse task slots. It is also used for medical image analysis because the modern imaging techniques like 3D and 4D result in a large file size which requires a large amount of storage and network bandwidth. Word count application is a common example of MapReduce.

4. GOOGLE FILE SYSTEM

Google File System is a closed source software. It is developed by Google. The codename of its new version is Colossus. Its interface is a familiar file system interface. Since there are hundreds of servers in a GSF cluster, a few of them could be

unavailable at any given time. To keep the system functioning, fast recovery and replication are the two strategies that are used. In fast recovery, the master and the chunkserver restore their state and start immediately irrespective of the way through which they were terminated. Replication is of two types: chunk replication and master replication. In chunk replication, every chunk is replicated on many chunkservers on different racks. In master replication, the master scale is replicated for dependability[10].

5. TOP 10 ALGORITHMS IN DATA MINING[11]

A. *C4.5*

C4.5 generates decision trees. These decision trees can be used for classification. They are also referred to as statistical classifier.

B. *The k-means algorithm*

The k-means algorithm is used to partition a given set of data into k cluster. It is an iterative procedure. The value of the number of clusters k is decided by the user.

C. *Support vector machines*

Support vector machines cope well with errors during execution. They are algorithms that perform two operations: Analyse data and recognize patters. There are two primary uses of support vector machines: Classification and estimating relationships among variables.

D. *The Apriori algorithm*

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

E. *The EM algorithm*

Em stands for expectation–maximization and it is an iterative algorithm. It is used to figure out maximum likelihood in models which are statistical and depend on variable that are not directly observed.

F. *PageRank*

PageRank is a search ranking algorithm using hyperlinks on the Web.

G. *AdaBoost*

The AdaBoost algorithm is used to group items to solve a problem or view them as a whole instead of single entities. It is very simple and has a great ability to make accurate predictions.

H. *kNN: k-nearest neighbor classification*

kNN is used for classification and regression. A label is classified to an unlabelled vector. The most recurrent among the k training samples, where k is the user-defined number, nearest to that query point is assigned the label.

I. *Naive Bayes*

If the probability an event is given i.e. the even has already occurred, then we can find out the probability of an event using the Bayes' Theorem.

J. *CART: Classification and Regression Trees*

CART is a method which is used for classification and regression trees. They are used for predicting variables that are dependent i.e. regression and predicting categorical variables i.e. classification.

LITERATURE SURVEY

Sr. No	Year	Paper	Description
1	1998	The PageRank Citation Ranking: Bringing Order to the Web	Discusses PageRank which is the method for rating web pages objectively and mechanically.
2	2001	Random forests	Shows that random forests are an effective tool in prediction.
3	2003	The Google File System	Presents the fact that The Google File System demonstrates the qualities essential for supporting large-scale data processing workloads on commodity hardware.
4	2004	MapReduce: Simplified Data Processing on Large Clusters	Shows that the MapReduce programming model has been successfully used at Google for many different purposes.
5	2004	A storage architecture for early discard in interactive search	Examines the concept of early discard for interactive search of unindexed data.
6	2005	Interpreting the data: Parallel analysis with Sawzall	Presents a system for automating such analyses.
7	2006	Integrating compression and execution in column oriented database systems	Discusses how to extend C-Store, which is a column-oriented DBMS, with a compression sub-system.
8	2006	Bigtable: A Distributed Storage System for Structured Data	Describes the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format along with the design and implementation of Bigtable.
9	2007	Dynamo: Amazon's Highly Available Key-value Store	Presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon's core services use to provide an "always-on" experience.
10	2007	An engineering perspective	Describes selected algorithmic and engineering problems encountered, and the solutions found for them.
11	2008	Top 10 algorithms in data mining	Presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006.
12	2009	"http://www.readwriteweb.com/enterprise/2009/02/is-therelational-database-doomed.php. ", "ReadWrite. "	Discusses the problem with Relational Databases.
13	2010	Google Big Table	Describes the differences between Bigtable and relational database and focus on the different data models used by them.
14	2012	A Few Useful Things to Know about Machine Learning	Summarizes twelve key lessons that machine learning researchers and practitioners have learned.

CONCLUSION

Bigtable provides good performance and high availability. The MapReduce model is easy to use, problems are easily expressible and its implementation scales to large clusters of machines. The technology that enables a very powerful general purpose cluster is Global File System. Thus the basics of Bigtable, MapReduce and Google File System were discussed in brief.

REFERENCE

1. *en.wikipedia.org/wiki/Data_science*
2. *en.wikipedia.org/wiki/BigTable*
3. *en.wikipedia.org/wiki/MapReduce*
4. *en.wikipedia.org/wiki/Google_File_System*
5. *Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. "Bigtable: A Distributed Storage System for Structured Data", 2006.*
6. *Xiao Chen. "Google Big Table", 2010.*
7. *Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters", 2004.*
8. *Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. "The Google File System", 2003.*
9. *http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/*
10. *http://computer.howstuffworks.com/internet/basics/google-file-system.htm*
11. *XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. "Top 10 algorithms in data mining", 2008.*