Big Data: Tools and Applications

Sofiya Mujawar
B.E. (last year)
Computer Engineering in MKSSS' Cummins college of Engineering, Pune

Soha Kulkarni
B.E. (last year)
Computer Engineering in MKSSS' Cummins college of Engineering, Pune

ABSTRACT
This paper states the study of some new and useful Big Data analyzing tools. Their architecture and applications are also mentioned. The various applications of big data in various fields today have also been discussed. An overview of both tools and applications have been comprehensively discussed.

General Terms
Machine algorithm, analysis database, relational database.

Keywords
Big data, big data analysis, tools, applications.

1. INTRODUCTION
The world today produces enormous amount of data every day. Intellects have also predicted that this scenario may also result in great wave of data or dramatically, even a data tsunami. This huge amount of data is now-a-days known as Big Data. More or less of the data tsunami being true, we now feel it a necessity to have a tool to have this data in a systematic manner for applications in various fields including government, scientific research, industry, etc. This will help in a proper study, storage and processing of the same.

2. CONCEPT
2.1 What is Big Data?
Big data is a term for large and complex unprocessed data. This data is difficult and also time consuming to process using the traditional processing methodologies.

Big data can be characterized as:

- Volume – The quantity of data that is generated is very important.
- Variety - Variety is the category to which Big Data belongs to is also a very essential fact that needs to be known for data analysis.
- Velocity - The term 'velocity' in the context refers to the speed of data generation or how fast the data is generated and processed.
- Variability - This is factor refers to the inconsistency which can be shown by the data at times. This can hamper the process of being able to handle and manage the data effectively.
- Veracity - The quality of the data being captured can vary to a great extent and hence does the accuracy.
- Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources.[1]

Thus to process this data, big data tools are used, which analyze the data and process it according to the need.

2.2 Goal of Big Data Tools:
Big Data tools are used for the analysis of the huge and complex data. Many organizations have now taken Big Data not just a buzz-word but a new technique for improving business.

Organizations have to analyze mixed structured, semi-structured or unstructured data. This is dons in search of useful business and market information and insights. Big data analytics helps organize this data for the organizations.

Organizations have to analyze mixed structured, semi-structured or unstructured data. This is dons in search of useful business and market information and insights. Big data analytics helps organize this data for the organizations.

Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information.

The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.[2]

2.3 Example
MongoDB

MongoDB [3] is a schema less document-oriented database. It is one of the prominently used tools for the analysis of unstructured data. Flexible to fit almost any use case. It is released under a combination of the GNU Affero General Public License and the Apache License. MongoDB is open-source and free software.

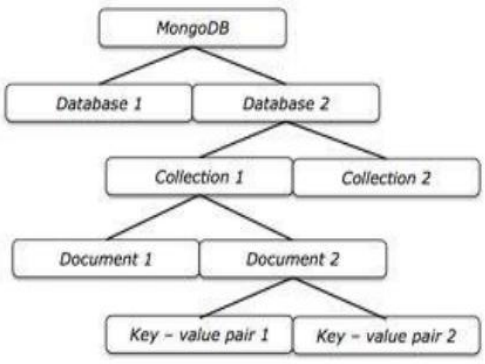

Fig1: Structure of MongoDB

3. BIG DATA TOOLS
The great amount of data collected can be classified into useful trends and patterns. Thus it must be preserved, studied and processed. Following are some of the majorly used Big Data taming tools:

3.1 Hadoop

Hadoop is a popularly used open-source data analysis tool. It is implementation of MapReduce for the analysis of large datasets. Hadoop uses a distributed user-level file system, to manage storage resources across the cluster. The file system is called HDFS, and is written in Java. It is designed for portability across heterogeneous hardware and software platforms.

Hadoop runs on the MapReduce model. In this, computation is divided into a map function and a reduce function. The map function takes a key/value pair and produces one or more intermediate key/value pairs. The reduce function then takes these intermediate key/value pairs and merges all values corresponding to a single key.[5]

The below diagram explain in brief the architecture of a Hadoop cluster. The client has transactions with the cluster. The cluster consists of cluster machines. Each of these cluster machines comprises of MapReduce agent and HDFS node. The cluster will also have a name node.

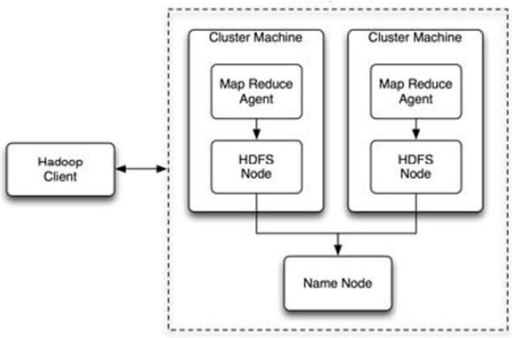

Fig2: Hadoop cluster

3.2 Google charts

The Google charts is basically an API tool. It is a free software. It lets people easily create a chart from some data and embed it in a web page. Google creates a PNG image of the required chart from data and formats parameters in an HTTP request.

It supports line, bar, pie, and radar charts. Also Venn diagrams, scatter plots, maps, Google-o-meters, and OR codes are supported.

For example, data about the oceans is provided. The Google charts tool will convert the data into simple diagram format like the one below.

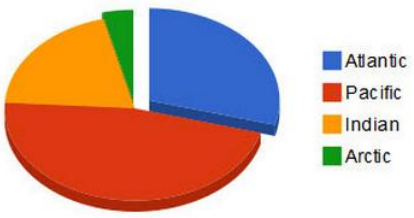

Fig3: Pie chart for oceans on the earth.

3.3 SAP's HANA

SAP HANA Enterprise 1.0 is an in-memory computing appliance that combines SAP database software with pre-tuned server, storage, and networking hardware from one of several SAP hardware partners.[6] It supports real-time analytic and transactional processing.

The distinctive features of HANA include:

- SAP's in-memory computing studio
- Sybase Replica Server 15
- SAP Host Agent 7.2
- Runs the SUSE Linux Enterprise Server 11 SP1 operating system

The indexer here will perform session management, authorization, transaction management and command processing. HANA has both row store and column store. One can create tables using either of the stores, but the column store has more capabilities. The index server also manages persistence between cached memory images of database objects, log files and permanent storage files.[7]

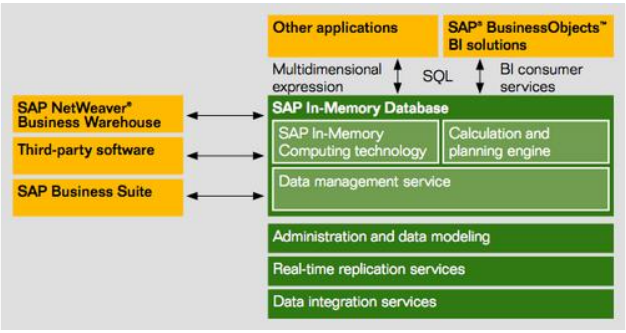

Fig4: SAP HANA Architectural Design
Source: SAP 2011

3.4 GridGain

GridGain is the leading provider of the open source In-Memory Data Fabric. It offers the most comprehensive in-memory computing solution. It helps equip the real-time enterprise with a new level of computing power. It enables high-performance transactions, real-time streaming and ultra-fast analytics in a single, highly scalable data access and processing layer. GridGain enables customers to predict and innovate ahead of market changes.

GridGain architecture can be explained in the below diagram. The GridGain In-Memory Data Fabric provides a unified API that spans all key types of applications like Java, .NET or C++, and connects them with multiple data stores. These contain structured, semi-structured and unstructured data (SQL, NoSQL, and Hadoop). This offers a very secure, highly available and manageable data environment, thus allowing companies to process full ACID transactions and generate valuable insights from real-time, interactive and batch queries.

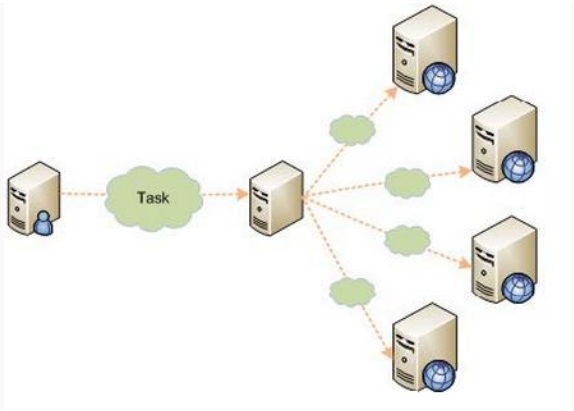

Fig5: GridGain architecture

3.5 Skytree

Skytree offers a package that can perform many of the sophisticated machine-learning algorithms. This requires the knowledge of the right commands. Skytree Server is optimized to run a number of classic machine-learning algorithms on your data. The implementation of these algorithms may help achieve a speed of about 10,000 times faster than other packages. It can search through the data looking for clusters of mathematically similar items, invert this to identify outliers.

Skytree also has its free version. It consists of the same algorithms. The major difference between paid and free version is that free version has limited to data sets of 100,000 rows.

Data from various resources is fetched firstly. This data is then transformed into the required format. It then goes for further processes to the Skytree.

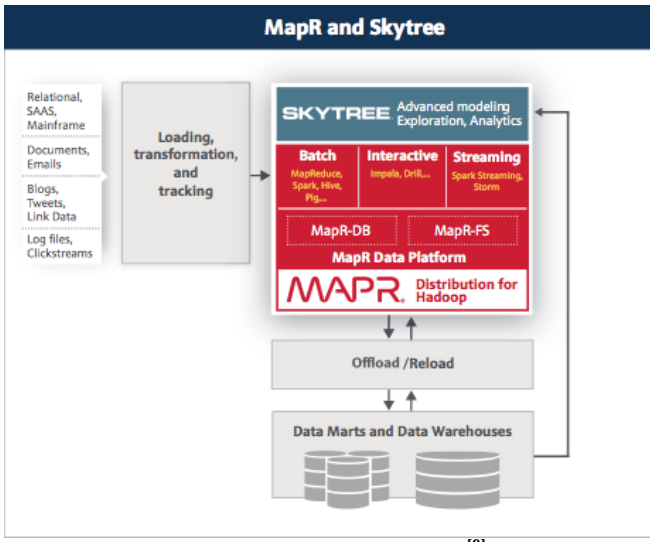

Fig6: Skytree architecture[8]

3.6 Splunk

Splunk is another analytics tool. It creates an index of the data as if the data was a book or a block of text. Although databases also build indices, Splunk's approach resembles more to a text search process. This indexing is highly flexible. Splunk tool is already tuned to a particular application, making it easier to make out the log files. The index helps correlate the data in these and several other common server-side scenarios.

Splunk will take text strings and search around in the index. Splunk finds the URLs one wish to find and packages them into a timeline built around the time stamps it discovers in the data.

The Splunk software architecture is explained in the below diagram. The data is fetched from the web servers and carried to the Splunk tool. This processed data then is transferred to the analytics database. The analyzed data is then transported to OLAP engine.

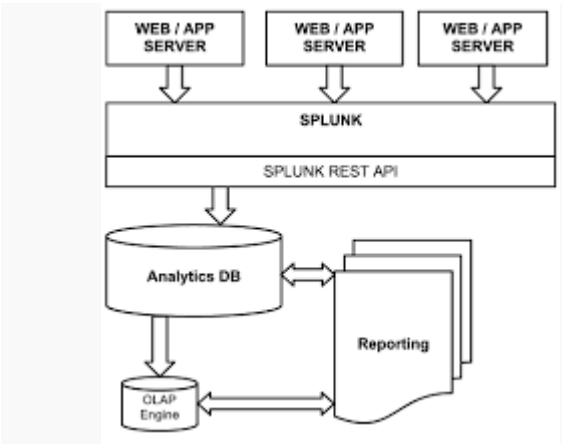

Fig7: Splunk architecture

3.7 Jaspersoft BI suite

Jaspersoft package is one of the open source software used for producing reports from database columns. It is one of the most leading software for Business Intelligence. This software is well-polished and is used to turn SQL tables into PDFs for better examination of data. JasperReports Serve offers software to take up data from major storage platforms, namely Mongo, Cassandra, Redis, Riak, CouchDB, and Neo4j. Jaspersoft not just offers particularly new ways to look at the data, but more of sophisticated ways to access data stored in new locations.

Once data is retrieved from these sources Jaspersoft converts them into lucid tables and graphs, thus making complex stuff easier. The reports are quite sophisticated and interactive helping one drill down into various aspects of it.

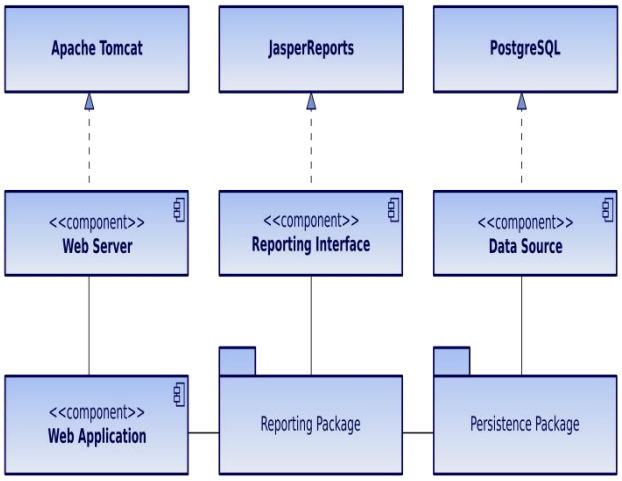

Fig8: Jaspersoft Architecture

4. BIG DATA APPLICATIONS

Big data has found many applications in various fields today. The major fields where big data is being used are as follows.

4.1 Government

Big data analytics has proven to be very useful in the government sector. Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign. Also most recently, Big data analysis was majorly responsible for the BJP and its allies to win a highly successful Indian General Election 2014.The Indian Government utilizes numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

4.2 Social Media Analytics

The advent of social media has led to an outburst of big data. Various solutions have been built in order to analyze social media activity like IBM's Cognos Consumer Insights, a point solution running on IBM's BigInsights Big Data platform, can make sense of the chatter. Social media can provide valuable real-time insights into how the market is responding to products and campaigns. With the help of these insights the companies can adjust their pricing, promotion, and campaign placements accordingly. Before utilizing the big data there needs to be some preprocessing to be done on the big data in order to derive some intelligent and valuable results. The ultimate goal of a company is to serve or convey a message or a product keeping in mind the consumer's mindset. Thus to know the consumer mindset the application of intelligent decisions derived from big data is necessary.

4.3 Technology

The technological applications of big data comprise of the following companies which deal with huge amounts of data every day and put them to use for business decisions as well. For example eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Facebook handles 50 billion photos from its user base. Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress. Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day. [10]

4.4 Science and Research

When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it collected data in its first few weeks which was much greater than all data collected in the history of astronomy. With a rate of 200 GB per night, SDSS has amassed more than 140 terabytes of information. Also its successor anticipated to acquire that amount of data every five days. Another example would be the decoding of the human genome which originally took 10 years to process, now it can be achieved in less than a day: the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law. The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[10]

4.5 Fraud detection

For businesses whose operations involve any type of claims or transaction processing, fraud detection is one of the most compelling Big Data application examples. Historically, fraud detection on the fly has proven an elusive goal. In most cases, fraud is discovered long after the fact, at which point the damage has been done and all that's left is to minimize the harm and adjust policies to prevent it from happening again. Big Data platforms that can analyze claims and transactions in real time, identifying large-scale patterns across many transactions or detecting anomalous behavior from an individual user, can change the fraud detection game.

4.6 IT log Analytics

IT solutions and IT departments generate an enormous quantity of logs and trace data. In the absence of a Big Data solution, much of this data must go unexamined: organizations simply don't have the manpower or resource to churn through all that information by hand, let alone in real time. With a Big Data solution in place, however, those logs and trace data can be put to good use. Within this list of Big Data application examples, IT log analytics is the most broadly applicable. Any organization with a large IT department will benefit from the ability to quickly identify large-scale patterns to help in diagnosing and preventing problems. Similarly, any organization with a large IT department will appreciate the ability to identify incremental performance optimization opportunities.

4.7 Call Center Analytics

Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful. What's going on in a customer's call center is often a great barometer and influencer of market sentiment, but without a Big Data solution, much of the insight that a call center can provide will be overlooked or discovered too late. Big Data solutions can help identify recurring problems or customer and staff behavior patterns on the fly not only by making sense of time/quality resolution metrics, but also by capturing and processing call content itself.

5. CONCLUSION AND FUTURE WORK

Handling big data efficiently is the need of the hour and one needs to come up with plausible solutions to these challenges one needs to understand the concept of big data, its handling methodologies and furthermore improve the approaches in analyzing big data. With the advent of social media the need for handling big data has increased monumentally. Approximately 5 Exabytes of data has been created, from the beginning of time till 2003. The same amount is now generated every 2 days. As more and more organizations are stepping out of the traditional boundaries big data keeps growing bigger. The tools being developed are efforts for overcoming the challenges arising due to big data.

6. REFERENCES

[1] http://en.wikipedia.org/wiki/Big_data#Definition

[2] http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

[3] Jing Han, Haihong E, Guan Le, "Survey on NoSQL Database" , IEEE 978-1-4577-0208-2,2011

[4] Sanobar Khan, Prof.Vanita Mane, "SQL Support over MongoDB using Metadata", ISSN 2250-3153

[5] Jeffrey Shafer, Scott Rixner, and Alan L. Cox , "The Hadoop Distributed Filesystem: Balancing Portability and Performance"

[6] Jeff Kelly, " Primer on SAP HANA".

[7] http://en.wikipedia.org/wiki/SAP_HANA#Architecture

[8] https://www.mapr.com/sites/default/files/otherpageimages/skytree%20diagram.png

[9] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[10] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[11] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[12] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[13] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[14] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[15] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[16] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[17] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender.

[18] http://en.wikipedia.org/wiki/Big_data