Big Data Analytics using Hadoop

Bijesh Dhyani
Graphic Era Hill University,
Dehradun

Anurag Barthwal
Graphic Era Hill University,
Dehradun

ABSTRACT

This paper is an effort to present the basic understanding of BIG DATA is and it’s usefulness to an organization from the performance perspective. Along-with the introduction of BIG DATA, the important parameters and attributes that make this emerging concept attractive to organizations has been highlighted. The paper also evaluates the difference in the challenges faced by a small organization as compared to a medium or large scale operation and therefore the differences in their approach and treatment of BIG DATA. A number of application examples of implementation of BIG DATA across industries varying in strategy, product and processes have been presented. The second part of the paper deals with the technology aspects of BIG DATA for it’s implementation in organizations. Since HADOOP has emerged as a popular tool for BIG DATA implementation, the paper deals with the overall architecture of HADOOP alongwith the details of it’s various components. Further each of the components of the architecture has been taken up and described in detail.

Keywords

Big data, hadoop, analytic databases, analytic applications.

1. INTRODUCTION

Companies across the world have been using data since a long time to help them take better decisions in order to enhance their performances. It is the first decade of the 21st century that actually showcased a rapid shift in the availability of data and it’s applicability for improving the overall effectiveness of the organization. This change that was to revolutionize the use of data brought into advent the concept that became popular as Big Data [1].

What is Big Data: Big data is the availability of a large amount of data which becomes difficult to store, process and mine using a traditional database primarily because of the data available is large, complex, unstructured and rapidly changing [2]. This is probably one of the important reasons why the concept of Big data was first embraced by online firms like Google, eBay, Facebook, Linkedin etc.

Big Data in small v/s big companies: There is a specific reason as to why Big data was first appreciated by the online firms and start-ups as mentioned above. These companies were built around the concept of using rapidly changing data and did not probably face the challenge of integrating the new and unstructured data with the already available ones [3]. If we look at the challenges regarding Big data being faced by the online firms and the start-ups we can highlight the following:

i. Volume: The largeness of the data available made it a challenge as it was neither possible nor efficient to handle such a large volume of data using traditional databases.

ii. Variety: As compared to the earlier versions, where data was available in one or two forms (possibly text and tables), the current versions would mean data being available additionally in the form of pictures, videos, tweets etc.

iii. Velocity: Increasing use of the online space meant that the data that was available was rapidly changing and therefore had to be made available and used at the right time to be effective [4].

1.1 The Challenges For Big Firms

Big data may be new for startups and for online firms, but many large firms view it as something they have been wrestling with for a while. Some managers appreciate the innovative nature of big data, but more find it “business as usual” or part of a continuing evolution toward more data. They have been adding new forms of data to their systems and models for many years, and don’t see anything revolutionary about big data [5]. Put another way, many were pursuing big data before big data was big.

When these managers in large firms are impressed by big data, it’s not the “bigness” that impresses them. Instead it’s one of three other aspects of big data: the lack of structure, the opportunities presented, and low cost of the technologies involved. This is consistent with the results from a survey of more than fifty large companies by New Vantage Partners in 2012 [6]. It found, according to the survey summary:

1.2 It’s About Variety, Not Volume

The survey indicates companies are focused on the variety of data, not its volume, both today and in three years. The most important goal and potential reward of Big Data initiatives is the ability to analyze diverse data sources.

Application areas and implementation examples:

1. Big Data for cost reduction: Some organizations that are pursuing Big data believe strongly that for the storage of large data that is structured, Big data technologies like Hadoop clusters are very cost effective solutions that can be efficiently utilized for cost reduction [7].

One company’s cost comparison, for example, estimated that the cost of storing one terabyte for a year was $37,000 for a traditional relational database, $5,000 for a database appliance, and only $2,000 for a Hadoop cluster.1 Of course, these figures are not directly comparable, in that the more traditional technologies may be somewhat more reliable and easily managed. Data security approaches, for example, are not yet fully developed in the Hadoop cluster environment [8].

1.3 Big Data At Ups

UPS is no stranger to big data, having begun to capture and track a variety of package movements and transactions as early as the 1980s. The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores over 16 peta-bytes of data [9].

Much of its recently acquired big data, however, comes from telematic sensors in over 46,000 vehicles. The data on UPS package cars (trucks), for example, includes their speed, direction, braking, and drive train performance [10]. The data is not only used to monitor daily performance, but to drive a major redesign of UPS drivers’ route structures. This initiative, called ORION (On-Road Integrated Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time. The project has already led to savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes [11]. UPS estimates that saving only one daily mile driven per driver saves the company $30 million, so the overall dollar savings are substantial. The company is also attempting to use data and analytics to optimize the efficiency of its 2000 aircraft flights per day [12].

1.4 Big Data For Time Reduction

The second common objective of big data technologies and solutions is time reduction. Macy's merchandise pricing optimization application provides a classic example of reducing the cycle time for complex and large-scale analytical calculations from hours or even days to minutes or seconds [13]. The department store chain has been able to reduce the time to optimize pricing of its 73 million items for sale from over 27 hours to just over 1 hour. Described by some as "big data analytics," this capability set obviously makes it possible for Macy's to re-price items much more frequently to adapt to changing conditions in the retail marketplace. This big data analytics application takes data out of a Hadoop cluster and puts it into other parallel computing and in-memory software architectures [14]. Macy's also says it achieved 70% hardware cost reductions. Kerem Tomak, VP of Analytics at Macys.com, is using similar approaches to time reduction for marketing offers to Macy's customers. He notes that the company can run a lot more models with this time savings.

1.5 Big Data For New Offerings

An organization can use big data to develop new products and offerings to it's customers. This is particularly effective for organizations that use the online space for it's products and services. With access to a large amount of data and the real time need of the customers, an organization can not only add value to it's existing offerings but also develop new offerings to match the needs of it's customers.

The best example may be LinkedIn, which has used big data and data scientists to develop a broad array of product offerings and features, including People You May Know, Groups You May Like, Jobs You May Be Interested In, Who's Viewed My Profile, and several others. These offerings have brought millions of new customers to LinkedIn.

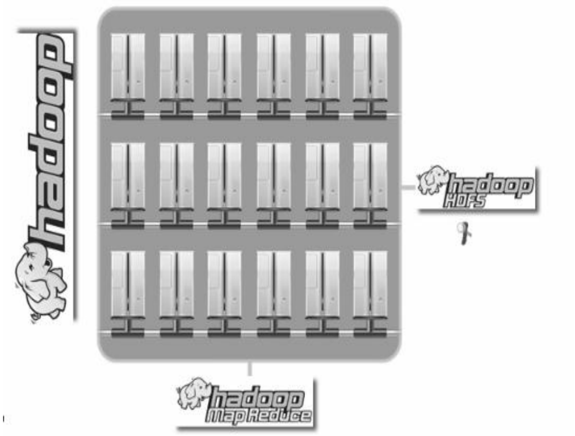

Fig 1: The figure shows the cluster where data can be inserted or capped.

Another strong contender for the best at developing products and services based on big data is Google. This company, of course, uses big data to refine its core search and ad-serving algorithms. Google is constantly developing new products and services that have big data algorithms for search or ad placement at the core, including Gmail, Google Plus, Google Apps, etc. Google even describes the self-driving car as a big data applications. Some of these product developments pay off, and some are discontinued, but there is no more prolific creator of such offerings than Google.

1.5.1 Example Of Big Data For New Offerings

Caesars (formerly Harrah's) Entertainment has long been a leader in the use of analytics, particularly in the area of customer loyalty, marketing, and service. Today, Caesars is augmenting these traditional analytics capabilities with some big data technologies and skills. The primary objective of exploring and implementing big data tools is to respond in real time for customer marketing and service.

For example, the company has data about its customers from its Total Rewards loyalty program, web clickstreams, and from real-time play in slot machines. It has traditionally used all those data sources to understand customers, but it has been difficult to integrate and act on them in real time, while the customer is still playing at a slot machine or in the resort.

In order to pursue this objective, Caesars has acquired both Hadoop clusters and open-source and commercial analytics software. It has also added some data scientists to its analytics group.

There are other goals for the big data capabilities as well. Caesars pays fanatical attention—typically through human observation—to ensuring that its most loyal customers don't wait in lines. With video analytics on big data tools, it may be able to employ more automated means for spotting service issues involving less frequent customers [15]. Caesars is also beginning to analyze mobile data, and is experimenting with targeted real-time offers to mobile devices.

1.6 Big Data For Improving Process Efficiency

Big data can be used for improving the process efficiency also. An excellent use of Big data in this regard is cricket specially with the advent of the Indian Premier League (IPL).

Not only are matches analysed using the data available in order to formulate future strategies but even minute details like the performance of a bowler against a particular batsman and that too on a particular ground under certain conditions are being made available for the stakeholders to improve their efficiency.

For example, how will a batsman like Glenn Maxwell perform against a bowler like Sunil Narine at Eden Gardens or how different will it be at Mohali in Chandigarh is available to be used. Not only this but also data like how many balls has a particular batsman faced against a particular bowler the number of dot balls and the number of runs scored [16].

Another example in this regard is the use of Big data to predict the probability (at any time during a match) of a team winning or losing in a match based on the extrapolation of the results in similar match situations..

2. HADOOP AS AN OPEN SOURCE TOOL FOR BIG DATA ANALYTICS

Hadoop is a distributed software solution. It is a scalable fault tolerant distributed system for data storage and processing.There are two main components in Hadoop :

(i) HDFS (which is a storage)

(ii)Map Reduce (which is retrieval and processing): So HDFS is high bandwidth cluster storage and it of great use what is happening here is (Fig. 1)

We put a pent byte file on our Hadoop cluster, HDFS is going to breakup into blocks and then distributed it to across all of the nodes of our cluster and on top of that we have a fault tolerant concept what is done here is HDFS is configure Replication Factor (which is by default set to 3). What does this mean we put our file on hadoop it is going to make sure that it has 3 copy of every block that make up that file spread across all the node in our cluster .It is very useful and important because if we lose a node it has a self feel what data was there on node and I am going to replicate that blocks that were on that node [17]. The question arise how it does that for this It has a name node and a data node generally one name node per cluster but essentially name node is a meta data server it just hold in memory the location of every block and every node and even if you have multiple rack setup it will know where block exist and what racks across the cluster inside in your network that's the secret behind HDFS and we get data.

Map Reduce: Now how we get data is through Map Reduce as name implies it is a two step process. There is a Mapper and Reducer programmers will write the mapper function which will go out and tell the cluster what data point we want to retrieve. The Reducer will then take all of the data and aggregate.

Hadoop is a batch processing here we are working on all the data on cluster, so we can say that Map Reduce is working on all of data inside our clusters. There is a myth that one need to be understand java to get completely out of clusters, infact the engineers of facebook built a subproject called HIVE which is sql interpreter. Facebook wants a lot of people to write adhoc jobs against their cluster and they are not forcing people to learn java that is why team of facebook has built HIVE, now anybody who is familiar with sql can pull out data from cluster [18].

Pig is another one built by yahoo, it's a high level data flow language to pull data out of clusters and now Pig and hive are under the Hadoop Map Reduce job submitted to cluster. This the beauty of open source framework people can built, add and community keeps on growing in Hadoop more technologies and projects are added into Hadoop ecosystem [19].

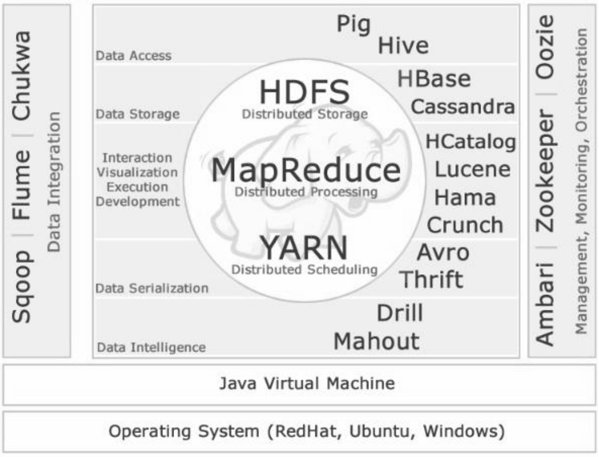

Figure 2. The image shows the hadoop technology stack. The hadoop core/common which consists of HDFS (Distributed Storage) which is a programmable interface to access stored data in cluster.

3. HADOOP TECHNOLOGY STACK

Figure 2. shows the hadoop technology stack. The hadoop core/common which consists of HDFS (Distributed Storage) which is a programmable interface to access stored data in cluster.

3.1 YARN (Yet another resource negotiation)

It's a Map Reduce version 2. This is future stuff. This is stuff which is currently alpha and yet to come. It is rewrite of Map Reduce 1.

3.2 Some essential Hadoop projects

Data Access: The reason why we need more way to access data inside Hadoop is because not everyone is low level, Java, C or C++ programmer that can write Map Reduce Jobs to get the data and even if you are something what we do in SQL like grouping, aggregating, joining which a challenging job for anybody even if you are a professional. So we got a some data access library. Pig is one among them. Pig is just a high level flow scripting language. It is really very very easy to learn and hangup. It does not have lot of keywords in it. It is getting a data, loading a data, filtering up, transforming the data and either returning and storing those results. There are 2 core components of PIG :

Pig Latin : which is a programming language

Pig Runtime : which competes pig latin and converts it into map reduce job to submit to cluster.

Hive : is another Data access project extremely popular like pig. Hive is a way to project structures on to data inside a cluster so it is really a database. Data-warehouse built on top of Hadoop and it contains a query language Hive QL like SQL to hive query language and it is extremely similar to sql. Hive is similar thing like pig. It converts these queries into map reduce jobs that gets submitted to cluster.

Moving Down to Technology stack we have:

Data storage: Remember out of the box is a batch processing system. We put the data into HDFS system; once we read in many time or what if we needed to get specific data; What if we want to do real time processing system on top of Hadoop data and that's why we have some of the column oriented database known as Hbase these are just Appache projects but there a buzz term for this NoSQL. Not once SQL that wants it stands for does not mean you can't use sql like language to get data out. What it means the underlying structure of the database are not strict like they are in relational world very loose, very flexible which makes them very scalable : that's what we need in the world of Big data and Hadoop, infact those are lot of NoSQL database platform out here. One of the most popular is Mangodb.

Mongodb is extremely very popular, especially among programmers because it is really very easy to work with. It is document style storage model which means programmers can take data models and clone. We call objects in those applications and serialize them right into Mongodb and with the same ease can bring them back into application [20][21].

Hbase was based on Google Big table, which is a way we can create table which contains millions of rows and we can put indexes on them and can do serious data analysis and Hbase is data analysis we put indexing on them and go to high performance which seeks to find data which we are looking for and nice thing about Hbase is pig and hive natively agree with Hbase. So you write pig and hive queries against data sitting in HBase table [23][24].

Cassandra is designed to handle large amount of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple data centres. It has its root in Amazon with more data storage table and it is designed for real time interactive transaction processing on top of our hadoop cluster. So both of them solve different problems but they both require seeking against our hadoop data.

We also have random collection of projects that span of different categories some of these solve specific business problems and some of them are little more generic like-

Hcatalog: It is known as metadata table and storage management system. What does it mean it's a way to create a shared schema, it's a way for tools like Pig, Hive for interoperable also to have consistent view of data across those tools.

Lucent: Lucent is there for full text searching an API loaded with algorithms to do things like standard full text searching, wild card searching, phrase searching, range searching kind of stuff.

Hama: Hama is there for BSP (Book Synchronous Processing). Here we need to work with large amount of scientific data.

Crunch: Crunch is there for writing and testing and running map reduce pipeline. It essentially gives you full control to overall four phrases, which is going to be: 1. Map, 2. Reduce, 3. Shuffle, and 4. Combine. It is there to help in joining and aggregation that is very hard to do in low level map reduce, so Crunch is there to make you little more easier inside map reduce pipeline.

Avro: These are data serialization technology which is a way in which we can take data from application, package up into a format that we can either store in our disk or send across the wires so that another application can unpack it and deserialize into a format that they can understand. Avro is more generic

Thrift is more specific for creating flexible schemas that work with hadoop data. Its specific because it is meant for cross-language compatibility, so we build an application with hadoop data in java and if we want to use same object inside an application that you built on Ruby, Python or C++.

Data Intelligence: We also have data intelligence in the form of Drill and Mahout.

Drill: Drill is actually an incubator project and is designed to do interactive analysis on nested data.

Mahout: Mahout is a machine learning library that concurs the three Cs :

1. Recommendation (Collaborative Filtering)

2. Clustering (which is a way to group related text and documents)

3. Classification (which is a way to categorize related text and documents).

So Amazon uses all this stuff to a further recommendation like music sites uses to recommend songs you listen and also to do predictive analysis.

Sqoop: On the left side of Figure 2. We have Sqoop. It is a widely popular project because it is easy to integrate hadoop with relational systems. For instance, we have result of map reduce. Rather than taking these results and putting them on HDFS, and require Pig and Hive query, we can send those results to relational world so that a data professional can do their own analysis and become the part of process. So, Sqoop is popular for pushing hadoop data into relational world, but it is also popular for pushing data from relational world into hadoop, like archiving [25].

Flume and Chukwa: are real time log processing tools so that we can set up our frame-work where our applications, operating systems, services like web-services that generate mountains of log data. It's a way we can push real time data information right into hadoop and we can also do real time analysis.

Over the right hand side of Figure 2., we have a tool for managing, monitoring and orchestrating all the things that go in our cluster:

Zoo Keeper: It is a distributed service coordinate. So it's a way in which we keep our all services running across all our cluster synchronous. So, it handles all synchronization and serialization. It also gives centralized management for these services.

Oozie: It is a work flow library that allows us to play, and to connect lot of those essential projects for instances, Pig, Hive and Sqoop.

Ambari: It allows you to provision a cluster which means that we can install services, so that we can pick Pig, Hive, Sqoop, Hbase and install it. It will go across all the nodes in cluster and also we can manage our services from one centralized location like starting up, stopping, reconfiguring and we can also monitor a lot of these projects from Ambari.

4. CONCLUSION

Doug Cutting, Cloudera's chief architect, helped create Apache Hadoop out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless..

5. REFERENCES

[1] M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.

[2] X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.

[3] Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE BigData, pp. 403-410, October 2013. A . Bellogín, I. Cantador, F. Díez, et al., "An empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37, January 2013.

[4] W. Zeng, M. S. Shang, Q. M. Zhang, et al., "Can Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation?," International Journal of Modern Physics C, vol. 21, no. 10, pp. 1217-1227, June 2010.

[5] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data," IEEE Trans. on Fuzzy Systems, vol. 20, no. 6, pp. 1130-1146, December 2012.

[6] Z. Liu, P. Li, Y. Zheng, et al., "Clustering to find exemplar terms for keyphrase extraction," in Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 257-266, May 2009.

[7] X. Liu, G. Huang, and H. Mei, "Discovering homogeneous web service community in the user-centric web environment," IEEE Trans. on Services Computing, vol. 2, no. 2, pp. 167-181, April-June 2009.

[8] K. Zielinnski, T. Szydlo, R. Szymacha, et al., "Adaptive soa solution stack," IEEE Trans. on Services Computing, vol. 5, no. 2, pp. 149-163, April-June 2012.

[9] F. Chang, J. Dean, S. mawat, et al., "Bigtable: A distributed storage system for structured data," ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39, June 2008.

[10] R. S. Sandeep, C. Vinay, S. M. Hemant, "Strength and Accuracy Analysis of Affix Removal Stemming Algorithms," International Journal of Computer Science and Information Technologies, vol. 4, no. 2, pp. 265-269, April 2013.

[11] V. Gupta, G. S. Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages," Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 2, pp. 157-161, May 2013.

A. Rodriguez, W. A. Chaovalitwongse, L. Zhe L, et al., "Master defect record retrieval using network-based feature association," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 3, pp. 319-329, October 2010.

[12] T. Niknam, E. Taherian Fard, N. Pourjafarian, et al., "An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering," Engineering Applications of Artificial Intelligence, vol. 24, no. 2, pp. 306-317, March 2011.

[13] M. J. Li, M. K. Ng, Y. M. Cheung, et al. "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters," IEEE Trans. on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1519-1534, November 2008.

[14] G. Thilagavathi, D. Srivaishnavi, N. Aparna, et al., "A Survey on Efficient Hierarchical Algorithm used in Clustering," International Journal of Engineering, vol. 2, no. 9, September 2013.

[15] C. Platzer, F. Rosenberg, and S. Dustdar, "Web service clustering using multidimensional angles as proximity measures," ACM Trans. on Internet Technology, vol. 9, no. 3, pp. 11:1-11:26, July, 2009.

[16] G. Adomavicius, and J. Zhang, "Stability of Recommendation Algorithms," ACM Trans. on Information Systems, vol. 30, no. 4, pp. 23:1-23:31, August 2012.

[17] J. Herlocker, J. A. Konstan, and J. Riedl, "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms," Information retrieval, vol. 5, no. 4, pp. 287-310, October 2002.

[18] Yamashita, H. Kawamura, and K. Suzuki, "Adaptive Fusion Method for User-based and Item-based Collaborative Filtering," Advances in Complex Systems, vol. 14, no. 2, pp. 133-149, May 2011.

[19] D. Julie, and K. A. Kumar, "Optimal Web Service Selection Scheme With Dynamic QoS Property Assignment," International Journal of Advanced Research In Technology, vol. 2, no. 2, pp. 69-75, May 2012.

[20] J. Wu, L. Chen, Y. Feng, et al., "Predicting quality of service for selection by neighborhood-based collaborative filtering," IEEE Trans. on Systems, Man, and Cybernetics: Systems, vol. 43, no. 2, pp. 428-439, March 2013.

[21] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 141-168, November 2005.

[22] Z. Zheng, H. Ma, M. R. Lyu, et al., "QoS-aware Web service recommendation by collaborative filtering," IEEE Trans. on Services Computing, vol. 4, no. 2, pp. 140-152, February 2011.