IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 130-136
www.iosrjournals.org

Big Data: The Future of Data Storage

Manpreet Singh, Jatinder Singh Bhatia, Devansh Malhotra
Associate Technology, Sapient Global Markets
Associate Quality Assurance Sapient Global Markets
Associate Technology, Sapient Global Markets

Abstract: *According to Internet World statistics, todayInternet has 1.7 Billion users, compared with the population of 6.7 billion people.Around 40% of the world population is connected via internet across the globe.By 2020 it is predicted, former will shoot to 5 billion users. Services are rapidly growing on the Internet, generating huge volume of Data likewise, trace logs, service information andtheir respective relationshipsetc. So, there is an urging need to gather such massive amount of Data. This is where the term-"Big Data"comes intopicture. Big data is a catch-phrase or a buzzword, used to describe a massive volume of both structured and unstructured data too large to be processed and handled using traditional database andsoftware technique. Considering, the volume of Data in context and its heterogeneity, its analysis is challenging.*
Focus of this paper is to do in depth survey as to what exactly is the upcoming technology -Big Data. ALiterature survey has been done to gather information of what has been done in the world of big data so far, present developments in it and the future of data storage under Big Data. The Data sources used for Big-Data analytics in the operational world simplydo not fit into desktop or small-scale database structures and therefore can be hosted using cloud concepts, to achieve this techniques are discussedas to how it is mined more efficiently, with on-premises database architectures.BigData is not a mere term, it's a concern now.

Keywords: *Big Data, Hadoop, Map Reduce, Cloud.*

I. Introduction

With increasing number of users demanding availability of services all the time, Big Data has become the driver for the innovation and Growth. Users these days want more personalized and more responsive applications. The vision of full time availability of the services and resources based on dataeverywhere, for anyone, at all time. BigData software and services generate value for innovative and intelligibleeco-system by enabling new-solutions. These values depend on advanced analysis.To accomplish this, we have various types of services of BigData for enhanced system performance. Big Data could be put to better use and availability by providing it as a service, rather than each member buying it. In context of services, different types of services available are - BigData Infrastructure Service (BDIAS), BigData Platform Service (BDPS), and BigData Analytics Software Service (BDSS), deployed to provide common

BigData related services to enhance efficiency and reduced cost for the users.Thus, Big Data has found its applicability in vast domains of technology.

Some of the interesting domains which may utilize Big Data include:

- Mobile systems and devices.
- Advanced Energy support systems.
- Personal health care systems and services.
- Transportation and logistics.
- Smart environmental systems and services;
- Intelligent systems and software engineering.

II. Background Analysis -

In 1970s, Cray Super computers had a monopoly over the processing and analysisof data. ENCAR,IBM, CRAY were some of the fastest among computers. These systems were too expensive and complicated for use w.r.t 1970swhere most of the companies could not afford such machines. Thesemachines used specialized memory & softwares but the configuration was not expendable; likewise, we couldn't take memory out of the CRAY and put to IBM. This was a draw back where load could not be divided, when one machine ran- out of memory or resources. Moreover, each had operating system with specialized programs and development methodologies. So, developers had to be well versed with all of them to code. This was redundant and less efficient. Additionally, more prone to human errors.

The concept of horizontal scaling and vertical scaling was used to conquer large amount of data.Super Computers of those times had limited vertical scaling. For Example- we could add some memory to the existing configuration but, once done we couldn't add another machine to balance the load. Another drawback of the technology was that processing of Data was not an easy nut to crack. It took a lot of time to process certain data as compared to Read, Write and transferring. But since, Processing was the largest part of the analytics. This was a tedious task to achieve.Therefore, various processing nodes or individual computers were used to process chunks of Data. So eventually, nothing came out to be in favor of the companies which drove companies away from the technology. Surprisingly, last decade saw a change.

A. Changes in technology over the Decade:-

Commodity Hardware available – Hardware is not an issue these days. Computers have well equipped with GBs Space for storage as well as processing. Memory Availability has become very cheap. Affording so much memory was too expensive earlier.

Today's computing benefits- Systems today are far more capable and fast to doing multi-processing and multi-tasking than ever before. The mobile devices which we carry in our pockets today were super computers of 70s era.

Commonly used languages – We have open source softwares and Internet Languages like JAVA, DotNet etc. are versatile and well known for development purposes. Previously such facilities were not available to the coders.

Network speeds are not the bottle necks - Networking was a big issue with low speeds. Thus allowing limited amount of sharing and processing.

Emerging storage needs of companies due to internet –

Companies these have days have vast appetite and requirement to store data. Social networking giants like Facebook store Petabytes of data. Amazon provides it Web services for storing data. Many organizations have their Hardware "Lying about". Google the famous search Engine has mammoth amount of data to store and process. The list goes on and so does the requirement of Big Data rising with each name enlisted.

B. Present Day Scenario - :

Upon entering the 21st century, the global economic structure has transferred from "industrial economy" to "service economy". According to the statistics of the World Bank, the output of modern service industry takes more than 60 percent of the world output, while the percentage in developed countries exceeds 70%. The competition in the area of modern service industry is becoming a focal point of the world's economy development. Service computing, which provides flexible computing architectures to support modern service industry, has emerged as a promising research area. With the prevalence of cloud computing, more and more modern services are deployed in cloud infrastructures to provide rich functionalities. The number of services and service users are increasing rapidly resulting in enormous explosion in Data generation by these services with the prevalence of mobile devices, user social networks, and large-scale service-oriented systems.

C. Cloud Beginning –

Cloud computing is logically related with Big Data. Cloud storage doesn't poses big issues though processing such Data with the concurrent conventional Data-mining algorithms are serious potential execution issues because of the anomalous and arcane structure data in context. Thus full usability of the cloud cannot be realized. Therefore, we need big data for such a scenario. But, cloud has its own benefits and also deltas to be covered. Listing a few:

D. Benefits of Cloud-

1. Achieved scalability of economy – Increased volume output or productivity with fewer people with economical cost per unit, project or product [12].
2. Minimal technology infrastructure cost– this maintains smooth access to the information with minimal upfront spending. Users are given flexibility in terms of payment ways.
3. Globalizeworkforce made affordable - People have access to the cloud across the globe, given that they have an Internet connection.
4. Multiple and stream Line processing - More work done in comparatively less time with less manpower.
5. Reduced Capital expenditures – Huge investment need not be done on hardware, software or licensing.
6. Quick Accessibility - You have easier faster and smoother access anytime, anywhere, making your life comfortable.
7. Projects effectively Managed–Stick to the budget and stay ahead in completion cycle times.
8. User friendly handling– Comparatively less personal training is needed, fewer people would do more work on a cloud, with a minimal learning curve relating to hardware and software bugs.
9. Improved flexibility and handling– Priorities could be assigned to the tasks, thus leading to affective management of work in hand. This is crucial for companies specially dealing with finances.

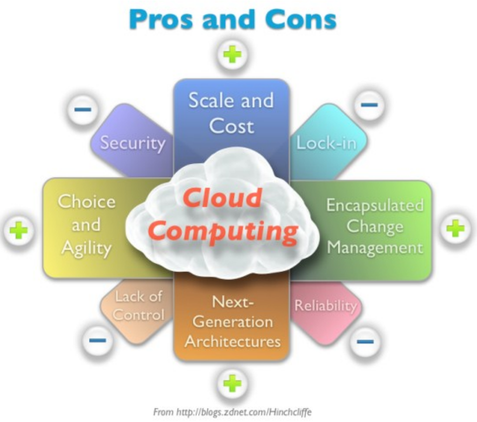
E. Mysteries of Cloud Computing-

1. Compatibility- Cloud is compatible with all the IT systems in a company. Albeit, cloud computing turns out to be one of the most cost efficient options for companies but, the problem arises out of the fact that the company might need to replace much of its existing IT infrastructures in order to make the systems compatible on the cloud.

2. Cloud Compliance–If the data of a company which is supposed to be "off the cloud", is stored on multiple servers, and sometimes distributed across several countries. This might lead to a trapin which a certain center develops an issue and cannot be accessed. Thus posing a serious problem for the involved company. Such a problem would be intensified if the data is stored in a server of a different country.
3. Standardization of Cloud - A contemporary problem linked with cloud computing is the current lack of standardization in the system. Since no proper standards for cloud have been set so far, it becomes almost impossible for a company to discern the quality of services they have been provided with.

4. Monitoring during usage ofthe Cloud-Monitoring of the cloudwhen taking services of a service provider is a problem. Once the all the authority is given to the provider, it might create monitoring issues for the user company.

5. Energy Requirement- The data for the cloud is stored in Data centers which consumes a lot of energy, thus we need a renewable transition.



From http://blogs.zdnet.com/Hinchcliffe

Fig.1 - Pros and cons of Cloud Computing

Present Day storage systems –
Relational Databases

The first and probably most obvious way of dealing with data is by using traditional Data warehousing architectures based on standard RDBMS(Relational Database Management systems). In this case, Data is extracted from various internal and external sources, selected, aggregated, and loaded into a Data warehouse. Different business intelligence tools can then be used to analyze and access the Data. As volume and velocity of the Data to be processed steadily increased since the 1980 [19], most contemporary companies revert to parallelized RDBMS to handle the large amounts of data [21]. Consequently, data are stored on multiple machines, tables are partitioned over the nodes in a cluster, and an application layer allows for accessing the different data portions on the different nodes. The goal of such an architecture is to provide linear speed-up as well as scale-up [21]. This means that twice as much hardware allows for execution of twice as large tasks in the same elapsed time. However, the effort necessary to keep the systems synchronized does not allow for completely linear scale-up or speed-up [22].

However, considering the RDBMS capabilities with regard to variety and velocity of data reveals several problems. As RDBMS are optimized towards data processing and analysis, the loading of data is typically very cumbersome and time-consuming.

III. Map Reduce And Distributed File System:

The second most often referenced strategy to approach data analysis is the introduction of new systems that use distributed file systems (DFS) and a MapReduce engine. A prominent exponent of such a system is Hadoop (although Hadoop does not exactly follow the MapReduce algorithm) [9]. Hadoop is an open source architecture composed of different engines such as a Map Reduce engine and a DFS engine. The Data to be analyzed is stored in the distributed file system and then processed using the Map Reduce engine. The results are then again stored in the file system and directly streamed to a business intelligence application. In the Map Reduce approach, unlike as in RDBMS, small programs are necessary to execute queries [25]. To be more precise, "users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key". These programs are thereafter, put into distributed processing framework, that decides how many map and reduce instances have to be run on which nodes [19].

Looking at opportunities and drawbacks of the MapReduce strategy, we come to an inverted picture compared to the RDBMS strategy. On the one hand, MapReduce lacks in its processing capabilities: the execution of standard tasks such as select and join can run up to 90 times slower on MapReduce systems than on parallel RDBMSs [19]. Additionally, it takes significantly longer to write a MapReduce program than an SQL query. Therefore, in environments in which large volumes of Data have to be frequently analyzed or in which analysis objectives (i.e. the underlying queries) change frequently, the time required for programming and waiting for analysis results hinders many organizations from implementing such a system [26].

A. Today's Problems in terms of Technology:

Technical problems include parallelizable [14] softwares (Open Sourcing) cannot be easily achieved. Also, optimum efficiency is to handle humongous amount of data is a cumbersome task. Another issue is Index building (Like Google). To build an index for huge undistributed data in itself is big task. Companies need softwares which could build indexes for their data to be processed and analyzed easily. Log Analysis is another maze which would mean that logging each and every detail in the records and maintaining the records for future instances. Human Problems also persists. For Example- Big Data exposes what we think is personal information likewise changing in buying products, choices of products for an individual. Another Challenge is disk I/O storing and accessing Data from disks. These problems can be dealt with following tactics.

B. Solution to the existing problems:

Parallelization- on technical front parallelization would mean to confront data parallel processing for every query relation to data access. MapReduce and Hadoop come handy in such a scenario, where data is Mapped and the processing time is reduced by parallel working from different nodes. Data that can analyzed in pieces. But, also there is a restriction to it; thing that do not work with Map Reduce would be any computational intensive task like Graph analysis or mathematical computation.Now, we want in some way to minimize the movement of data from center to the other nodes for processing. Solution to disk storage is using Stripping and mirroring concept where data is stripped first and simultaneously mirror for avoiding and loss or termination of data while processing.

IV. Big Data Framework

Architectural Design of the Cloud Based Big Data Analysis.

Our guiding design principle for the Cloud-based analysis service is to reuse existing, well-tested tools and techniques.

Architecture of Hadoop–

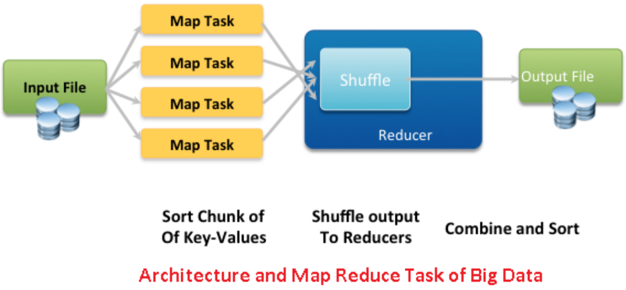


Fig. 2- Architecture and Map reduce task of Big Data

- Node - A node is a nothing but and individual computer
- Rack - Collection of nodes on a network.
- Network Bandwidth - Connectivity between racks and their respective nodes. Bandwidth between nodes within a rack is always greater than between different racks
- Cluster - Collection of different racks

A. Apache Hadoop:

Open source software framework for storage and processing of large scale-Data sets on cluster of commodity hardware [10]. It is built on java framework and uses Google's Map Reduce and Google file system. It does massive parallel processing done with great performance. Hadoop is not suitable for OLTP and OLAP which are accessed using the structured Data. Hadoop works more efficiently on larger files. Does sequential data access rather than random access. Hadoop is designed to work on the entire data set.

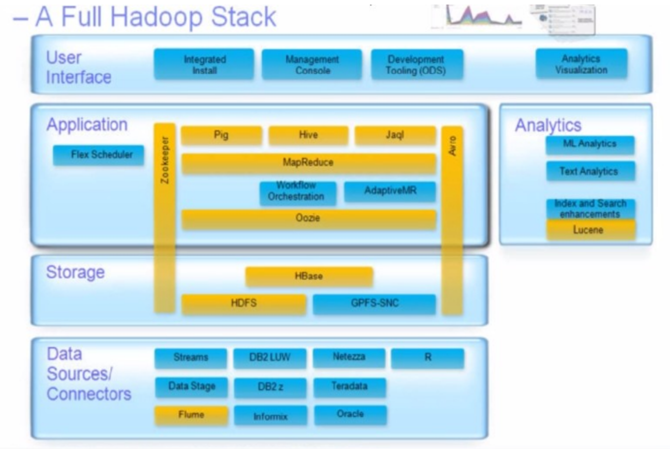


Fig.3 – Hadoop Full stack

B. Phases for analyzing Data:

Acquisition→Extraction→Integration→Analyzing→
Interpretation

C. Modules of Hadoop:

1. Hadoop Common- Libraries and utilities for other modules.
2. Hadoop Distributed File system- Provides very high aggregate cluster Bandwidth across the cluster by storing Data on commodity machines.
3. Hadoop YARN- Resource management system to manage compute resources in cluster to schedule user applications. It provides generic scheduling and resource management. Hence, this way Hadoop can support more than just Map Reduce[10]. It also provides work load management and efficient scheduling.
   With YARN the resource manager is already aware of the capabilities of each node via communication with node manager. Else this was required to be done with Map Reducer manually.
4. Hadoop Map Reduce – It is designed to process very large Data sets, for distributable problems. It allows the work to be spread across different nodes and make them work in parallel, provided there are no dependencies.

D. Map work and Reduce operations –

1. First Map the set
2. Takes the sub set of full- Data called an input split and applies to each row.
3. The Output is buffered to memory, sorted and partitioned by using default practitioner.
   Now, these partitions are sorted using Merge sort. These sorted partitions are sent to Reducers. Example- 1 Map task partition to reducer 1 and 2 Map task also sends its partition1 to reducer 1 and so on.
4. Each reducer further reduces tasks by its codes.

V. BigData Analytics as a Service:

The fourth strategy for dealing with BigData is to buy in BigData capabilities. The supply of BDAaaS solutions is rapidly increasing and the variety of vendors is large. Tresata, for instance, has specialized on analyzing banking Data [33]. Another example is Cloudera [17], who offers a large variety of BDAaaS solutions for different industries. Therefore, organizations tackling the BDAaaS strategy are likely to find suitable solutions for their specific context.

As the infrastructure for BDAaaS is hosted in the cloud, the costs for BDAaaS in organizations are far more flexible than the costs of implementing in- house BigData solutions. In addition, the economies of scale and scope realized at the side of the vendor allow them to perform BigData analytics more efficiently than an average company could do [34]. This is especially interesting for smaller organizations, which often do not have sufficient resources and expertise to realize BigData analytics.

However, BDAaaS has limitations as well. The most critical limitation arises from the Data privacy and security discussions that are relevant for all cloud-based services. Especially in operating environments in which sensitive Data has to be analyzed in the cloud, an encryption during the Data transmission is inevitable. Operating with encrypted Data however is yet a problem for most BDAaaS solutions. Further, Data policy is already an important issue, when BigData is analyzed using internally-secured infrastructure. Therefore, shifting such analyses into the cloud may be problematic for some companies, if not impossible. Besides, there is no definite answer, yet, on how to exchange Data between cloud providers and company-internal infrastructure [34]. As a result, many cloud providers offer BigData analytics only for company-external Data that can be retrieved without access to internal infrastructures

Problems that remain to be solved are as follow:

- Capacity - Capacity can reach the scale of PB/ZB level. As a result, mass Data storage systems should have ability for scaling. Meanwhile, the online expansion method must be convenient enough and decrease the degradation serving time;
- Delay: BigData applications should be real-time, especially involved with online trading or financial related applications. Date-intensive computing has the SLA (Service Level Agreement) requirements. In addition, the popularity of server virtualization has led to a stringent demand for the high IOPS;
- Security and Privacy: Access control for BigData is also a hot topic for research. BigData applications lead to concern about security, especially for private Data of company or individual
- Cost: To save the cost, we need to make every device work more efficiently and avoid over-provision. Due to the widely implementation of coding, Data de-duplication and cloud storage technologies in storage field, BigData storage applications can be more effective and valuable.

VI. Future Of Bigdata

- Companies are optimizing their work around BIGDATA
- The Executive mind set around fact based decision is changing in everything.
- Data Governance is coming into picture.
- Banks are starting to realize that how their customer are progressing, so that they can cover these divergent channels.
- To increase the predictive power of an organization
- The more company knows about its past, more it can approach the customers.

“We can and should use BIGDATA to “Understand today and design tomorrow”.

VII. Conclusion

Processinglargerdatasetshasbecomeincreasinglypossibleoverthepastfewyearsforamuchlargercommunity, notleastviathedevelopmentoftheMap-Reduceparadigm.Map-Reduceenablethepowerofparallelcomputingtobeavailabletostandarddataanalysistasks. Asmentionedbefore, themainchallengeinapplyingmanyislandsofBigDataapplicationsistoidentifythedefininglinesofeachapplicationand theirinter-relationship. Since Data processing and analyzing is increasing at an alarming speed, we need a change in future technology to be more efficient and versatile to enable users to access data at anytime and anywhere in their customized way. Inordertomanageandintegratethisspreadwithinanorganization.ToclarifywhatBigDatareallyis,itistheenterprisedata processingenvironmentforheterogeneousdataandcomputationalsourcesinatimelymannertogaincompetitiveadvantage.Thisresultsintheprocessingofhighvolumeofdataandpresentingthisinaconciseandclearmannertoaidoperationalandstrategicdecisionmaking.

References

[1] The Economist, Nov 2011, "Drowning in numbers – Digital Data will flood the planet and help us understand it better", http://www.economist.com/blogs/dailychart/2011/11/Big-Data-0

[2] 2013IEEEInternationalConferenceonSystems, Man, and Cybernetics,”Big Data Framework ”Firat Tekiner1 and John A. Keane

[3] Bonnet L., Laurent A., Sala M., Laurent B., Sicard N., September 2011, "Reduce, You Say: What NoSQL Can Do for Data Aggregation and BI in Large Repositories", dexa, pp.483-488, 22nd International Workshop on Database and Expert Systems Applications, 2011

[4] Mark B., "Gartner Says Solving 'BigData' Challenge Involves More Than Just Managing Volumes of Data". Gartner, June 27, 2011, http://www.gartner.com/newsroom/id/1731916

[5] 2013 IEEE International Symposium on Multimedia,‘BigData and NSA Surveillance - Survey of Technology and Legal Issues’ ,Chanmin Park Taehyung Wang

[6]] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, “Interactions with BigData analytics,” interactions, vol. 19, no. 3, pp. 50–59, May 2012

[7] N. Wingfield, “Virtual product, real profits: Players spend on zynga’s games, but quality turns some off,” Wall Street Journal.

[8] T. White, Hadoop: The Definitive Guide. ,2009

[9] 2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops,'Towards Service-oriented Enterprise Architectures for BigData Applications in the Cloud', Alfred Zimmermann, Michael Pretz

[10] International Journal of Computer Applications (0975 – 8887) Volume 80 – No.9, October 2013,"Trustworthiness of BigData"Akhil Mittal.

[11] http://cloudcomputing124.blogspot.in/2013/03/cloud-computing-benefits-both-financial.html#.U_LomPmSzkU

[12] http://www.verio.com/resource-center/articles/cloud-computing-benefits/

[13] http://www.forbes.com/sites/joemckendrick/2013/01/30/7-great-unsolved-mysteries-of-cloud-computing/

[14] http://techbus.safaribooksonline.com/video/databases/hadoop/9781771372374

[15] http://www.zdnet.com/blog/hinchcliffe/eight-ways-that-cloud-computing-will-change-business/488

[16] http://techbus.safaribooksonline.com/video/databases/hadoop/9781771372374