iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: IX Month of publication: September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38038

Transformation in Data Storing Technique- Big Data: A Literature Review

Adarsh Neema
Department of Computer Science, Medicaps University, Indore, India

Abstract: Loss of data implies loss of valuable information. An appropriate gathering of data and finding hidden patterns out of it is key for any business organization to thrive fiscally. With exponential rise in the internet users from the early 2000's, traditional databases fell short to collect the enormous amount of unstructured data/ semi-structured data, which contained extremely insightful information. Today, data accumulated is not only enormous, but also collected with high speed, having variety, which necessitated special database/software for data gathering and making key decisions based on that. These gigantic amounts of data generated can advocate companies to examine the market trends, market demands, and customer expectation, which endorses them to make relevant foremost decisions. This study discusses the stymie in conventional databases to process the immense data and entailment of advanced databases/software. Furthermore, a case study is presented later to understand the role of big data analytics in business and technical firms.

Keywords: Big Data, structured data, unstructured data, NoSQL, Hadoop.

I. INTRODUCTION TO DATA

When it comes to defining the term data, it does not have a single definition but can be expressed in disparate ways. Earlier it was believed, data is just related to spreadsheets, tables or charts that incorporate numbers and labels, but not just limited to this. In layman terms, data is defined as accumulation of factual information, observations, statistics or description of things that can be used as a reference to draw insights and make decisions.[1]

But the term data is not contemporary, it has been in use for centuries from antediluvian period just with different names. Even before the invention of computers, ancient people used distinct tools such as tally sticks, ishango bones, abacus etc. to store and keep track record of their activities and supplies and made calculations based upon that.

Before the rise of the internet, the amount of data was limited and therefore traditional databases such as SQL, MS-excel etc. were enough to store and analyze the data generated. Earlier the data was entirely in structure format with fixed schema which allowed the DBA(database manager) to swiftly store and manage data without any memory issues. Moreover, the number of internet users in early 2000's were just 361 million worldwide therefore, the velocity with which data was assimilated was minute.

With advancement in technology, the usage of the internet rose, meaning more generation of data. By end of year 2010, the number of internet users reached upto 1967 million which directly implies higher velocity of data gathering. With rise in volume and velocity of data, traditional databases fell short of storing and accumulating this data, not only because of size of data but also due to type of data. The data collected was largely either unstructured or semi structured, creating difficulties for traditional databases and softwares to store these data. There was a need for tools and technologies that could store these voluminous data with full pledge security and that could prevent data loss and hence this was the rise of the term big data.[2]

The term big data refers to the data that is so immense, quick and perplexing that it's onerous to process using traditional methods. special softwares and databases are designed to process such enormous data. Generally, the term big data is characterized by 5 v's namely- volume, velocity, variety, veracity and value.

1) *Volume*: As name specifies, volume refers to the amount of data that has been produced. As big data involves structured, unstructured and semi structured data, the size of data has been perpetually increasing. Within the last two years, the amount of data generated is so much that it comprises 90% of the world's data. The volume of big data will be ever increasing with the rise in internet users. This can be analyzed by the fact that just 2 zettabytes(10^21) of data was generated in 2010 and reached upto 41 zettabytes 2019, 59 zettabytes in 2020 and expected to increase upto 74 zettabytes in 2021 and 180 zettabytes in 2025.
2) *Velocity*: Velocity refers to the speed with which data is generated, collected and analyzed. With excessive use of devices and the internet, data flows from machines, social media, cell phones etc. at a higher rate which is further used for data visualization. Considering the volume of data accumulated, it can be concluded that from 2010 till 2021, speed of data collection has tremendously increased.

a) *Veracity*: Veracity refers to the quality of data. Even if the bulk of data is collected, it will not be useful if data is messy and inconsistent which will be extremely hard to analyze.

b) Value: Value of data defines how much data is actually useful and meaningful. The bulk of data generated would not be trivial if it cannot draw any insight and analyze the data.

c) *Variety*: Variety defines the nature of data that is flown. Today data collected is generally in three formsnamely-

- *Structured Data*- Meaning collected data follows proper schema which is organized and defined in systematic format. Today, structured data is less than 20% of total, implying most of the data gathered does not have anyform.
- *Unstructured Data*- Meaning the data that does have any format due to which it does not fit into any row and column and becomes hard to analyze. Today, at most 80% of data accumulated is in unstructured form
- *Semi-Structured Data*- Meaning the data which has structure but cannot be stored in traditional databases and does not follow tabular structure of data.

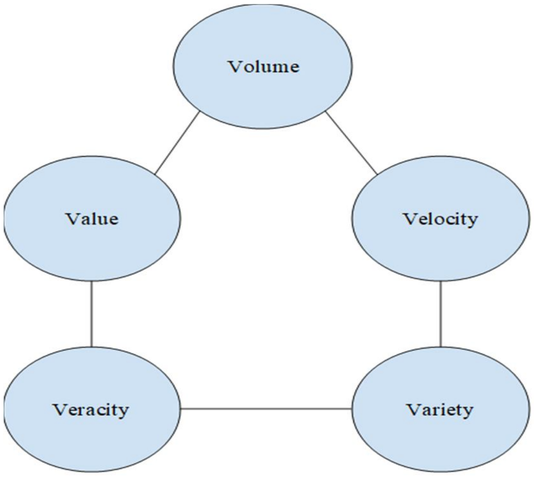

Fig 1. 5 v's of Big Data

II. NEED OF BIG DATA

As today, most of the businesses are shifting their base from physical mode to online mode, they are noticing ample amounts of data assembled in all formats. Companies are leveraging use of these data to give a cutting edge to their business and identify new opportunities. Mentioned below are few merits of big data:

A. Enhancing Customer Services

Big data focuses on customer satisfaction and prioritizing their needs, based upon their buying preference- what and when they purchase and makes similar recommendations within the same budget.

B. Analyzing Market Condition

Big data substantiates industries and organizations to keenly apprehend the market situation, which gives them competence to stay ahead of their adversary.

C. Innovation in Product Development

Big data helps the companies to keep an eye on their product in the market. They use the reviews collected to make innovation in the product and redevelop it’s best possible version. Moreover, big data helps them to understand which of their products saw a spike at what point of time and therefore, the companies are ready with ample products at that time period.

D. Risk Management

To stay from all the risks and Fraudulent activities is the primary focus of all the organizations. For this purpose, they leverage big data analytics to identify and narrow down any discrepancies, which helps them to find the root cause of the problem.[3]

III. EVOLUTION OF DATA STORAGE

Over the years, the methods and ways of storing the data have greatly changed. From calendar year 1995, where internet users were just 16 million, which was merely 0.4% of world population to calendar year 2021 where internet users are 5168 million which is 65.6% of world's population, we have come long way along, which directly reflected the 5 V's(volume, velocity, variety, veracity and value) of data. With proliferation in internet users, which caused 5 v's of data to thrive, there has been tremendous change in data storing techniques. Before early 2000's, when data generation rate was exiguous, traditional databases were enough to collect, organize and analyze the data, to draw inferences but from early 2000's onward, with rapid growth of technological world, 5 v's of data stared to increase at rapid pace which created catastrophe in storing, managing and analyzing the data. The inability of traditional databases to store huge chunks of unstructured data created a problem of data loss which directly implied loss of information. This couldn't be continued for a long time and required a novel system with ameliorated access performance which can efficiently handle these voluminous data with higher scalability and reliability, and hence this was how the term "Big Data" was introduced.[4]

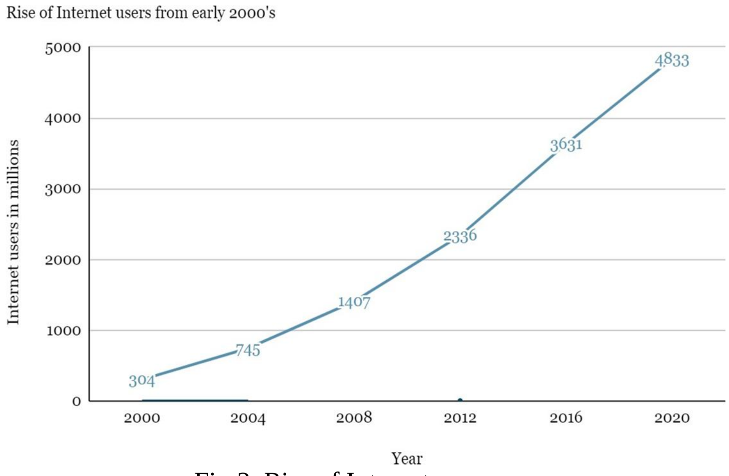

Fig 2. Rise of Internet users

Table1. Swot analysis of structured and unstructured data:[5]

PARAMETERS	STRUCTURED DATA	UNSTRUCTURED DATA
Strength	Due to fixed schema and simple organization, they are easily perceivable and manageable with easy data entry, search, comparison and extraction.	Due to its voluminous amount, it offers more profound insights meaning better decision making to discover gaps in the market and provide innovation.
weakness	Structured data is stored in a data warehouse, therefore any amendments in data requirements means to update the whole data warehouse leading to more time consumption and involutions.	Having an irregular schema, not everyone can work with unstructured data, therefore it requires profound expertise as a data scientist to understand unstructured data.
Opportunities	It provides a system to support involute queries with atomicity in transactions. Also, it can automate the software development process.	Unstructured data or big data has faculty to endorse it's benefits in every field, even with micro industries and provide opportunity for in-depth decision making.
Threats	Storage is the primary issue because of dynamic growth of data. Additionally, with predefined schema, it can be only used for defined purposes which limits its use.	As the volume of data is surpassing it's limit every day, there is a challenge with the quality of data, which may lead to ineffective decision making.

Big data has greatly discerned it's significance in forecasting the business results, but it came with setbacks such as data security, merging data from disparate sources, proper tools to store big data, preventing data loss etc. In order to summon the challenges of big data, NoSQL database(not only SQL) has been surfaced as a top-notch solution, which offers the faculty to gather and analyze structured, semi-structured and unstructured data. NoSQL collectively solved all the major issues associated with traditional databases and with improved scalability, flexibility, accessibility, prevented data loss by storing replicas of data, data security and fast processing. NoSQL stores data in distributed fashion and offers parallel processing designed chiefly for database frameworks.

On the same side, Hadoop has emerged as one best open source framework for storing unstructured data and running applications on the cluster of the commodity hardware which consists of HDFS(Hadoop file system) for massive parallel processing and MapReduce for data processing. HDFS is largely responsible for storing the big data and collecting the data on multiple chunks to prevent data loss, while MapReduce divides up the task among different mapper classes in the form of (key,value) pair and reducer class binds the output from the mapper class. Both NoSQL database and Hadoop software control gigantic and swiftly growing data with distinct data formats but depending upon the business and technical need, they may differ slightly.[6]

Table 2. Difference between NoSQL and Hadoop

PARAMETERS	NoSQL	Hadoop
Purpose	NoSQL is best suited for operational workload, which provides the ability to query the data.	Hadoop is best suited for analytics purposes, to generate insight from the data.
Limit	NoSQL can handle as large as terabytes of data.	Hadoop can handle as large as petabytes of data.
Time to process	As NoSQL works on small subsets, it takes seconds to process the data.	Hadoop works on large subsets, therefore it takes minutes or even hours to process the data.
Use	NoSQL is developed for real time interactive access to data.	Hadoop is developed for large scale data processing used for analysis.

IV. TYPES OF BIG DATA ANALYTICS

In today's era it is not to say but, without data, companies and business organizations wouldn't be able to achieve the level of success that they sought to accomplish. Today, revolution in technology has achieved its climax and has complete domination over our lives and our actions by pertinently gathering data. Appropriately assembled data, when sorted in acceptable and proper fashion can find out the course of action the company needs to follow in order to flourish. Disparate types of data necessitates different and unique approaches. This section discusses different types of big data analytics, along with real life example of how amazon, the multi-billionaire company like Amazon uses these different types of big data analytics to overcome the challenge and build better system-[7]

A. Descriptive Analysis

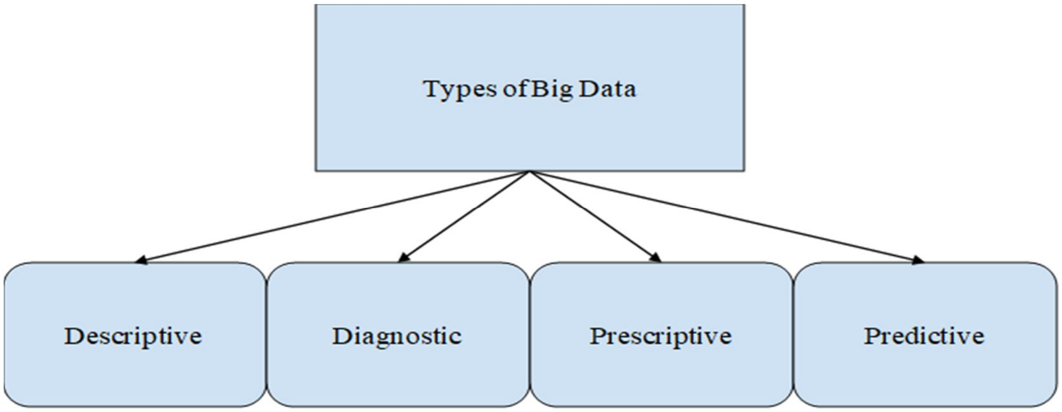

Fig 3. Types of Big Data

Descriptive analysis summarizes past data into a form that is interpretable by humans. It tries to answer the questions about the past trends by analyzing the past records. This analytics helps to create reports like company's revenue, sales, profit etc. and provide information of past trends that have been followed. For e.g.: Amazon uses descriptive analysis by creating annual reports of their sales, warehouses, workflow and KPI's by using graphs and visuals to get the insight about how it has been performing. Based on reports, the company makes various important decisions to minimize loss and to generate a large customer base.

B. Diagnostic Analysis

Diagnostic analysis is performed in order to understand why a particular problem has occurred. It is the cause of the problem. This analytics is characterized by techniques such as drill down, data mining and data discovery. For e.g.: Amazon uses diagnostic analysis to create reports and analyzes, what possibly has gone wrong ? If the company's sales have been degrading, even after the customer's adding product to their cart, then why are they not buying the product ? Was the shipping fee too high? Was there not enough payment options ? Or what else? And therefore uses diagnostic analysis to find answers to all these questions.

C. Predictive Analysis

Predictive analysis digs deep into the historical and present data to make predictions of the future. This analytics uses data mining, artificial intelligence and machine learning to analyze current data to make predictions about the future and works on probability. For e.g.: Amazon uses predictive analysis to determine what kind of precautions they have to take in order to protect their customers from fraudulent transactions. Using predictive analysis, Amazon uses all historical data, the user behavior data and builds an algorithm which predicts fraudulent activities.

D. Prescriptive Analysis

Prescriptive analysis prescribes a solution to a particular problem. This type of analytics is a blend of both, descriptive analysis and predictive analysis. Main purpose of prescriptive analysis is to endorse the best suitable solution for a problem from the available data. For e.g.: Amazon uses prescriptive analysis to maximize their profit. They build and use an algorithm that automatically adjusts the product rate based on numerous factors including but not limited to- customer demand, season, rating of the product etc.

V. TECHNOLOGIES AND BIG DATA

With the thriving succession of automation in technology innovation along with inclination in customers demand, the requirement of a systematic data management system has become supreme. Even minute information about customers and their purchasing and selling behaviors is the chief source of the database(unstructured data), which advocates the operation of big data for real time decision making. For generating more profits and to create a proliferating customer base, there is an utter need for the business organization to shift from physical mode to online mode to digitize their business. This has led to an extreme need for systems and technology, which can handle companies enormous data. With big data profoundly in use, disparate databases and software have been developed to ameliorate the functionality of big data. Technologies incorporated in Big data serve the purpose of data mining, data storing, data visualization to make an analysis of data which is of foremost importance for decision making. Some of the primarily used technologies in big data are discussed below:

A. Hadoop Ecosystem

Hadoop ecosystem is the open source framework, which has the capabilities to store data in parallel manner, with replicas of data stored in different machines to prevent data loss. When it comes to processing big data, the Hadoop framework is the most preferred one because of its ability to store as large as petabytes of data. Since Hadoop has been widely in use, large companies have been using Hadoop as a data warehouse to store and process their big data.

B. Artificial Intelligence(AI)

The use of artificial intelligence in big data has transformed the big data process. The functionality of AI to achieve a definitive goal without being explicitly programmed and without human intervention has made its existence beneficial. Because of artificial intelligence's ability to make predictions about the future, analyzing the current situation and redeveloping the business model with highest precision, the requirement for humans to manually perform these activities has been greatly declined. Today, AI is used in big data to advocate humans with daily basis actions and automating the process of decision making.

C. NoSQL Database

Not only SQL or NoSQL, was designed to process all three formats of data(Structured, Semi-Structured and unstructured data). NoSQL supports a large storage unit to process unstructured data with flexibility, Scalability, higher performance and high functionality. NoSQL imparts disparate big data technologies in their databases for data analyzing, decision making and application development, for instance- MongoDB, Cassandra etc. Currently, NoSQL is employed for real time data analysis, which holds the capacity to store terabytes of data.

D. Tableau

Tableau is a powerful and fastest growing data visualization tool, used for analysis of data by creating interactive dashboards and worksheets and sharing big data insights. Tableau uses the big data and converts it into interactive graphs and pictorial forms and bolster with relevant decision making.

E. Blockchain

Blockchain is a big data technology, where each record in the database is in the form of a block. Every new transaction that takes place creates a new block, and all these blocks are interrelated by chain, forming blockchain. Data once stored becomes immutable, therefore cannot be altered which makes it thoroughly secure. Blockchain when employed with big data can create a powerful system. Currently both of these technologies are used concurrently in secure, low-cost online transaction and data storage. Secondly, Blockchain technology can be useful in fraud detection by tracing the transaction to its origin to detect any aberration. And lastly, when both technologies used together can find patterns within the data such as customer habits.[8]

VI. LIFE CYCLE OF BIG DATA ANALYTICS

The data in big data passes through a series of steps, before processing and giving the final analysis of the result, in order to sort out and arrange the activities. The analysis differs from traditional data analytics due to 5 v's of Big data. This section discusses the life cycle of big data, depicting how data is utilized and analyzed in order to make business insights and solutions.[9]

A. Business Case Evaluation

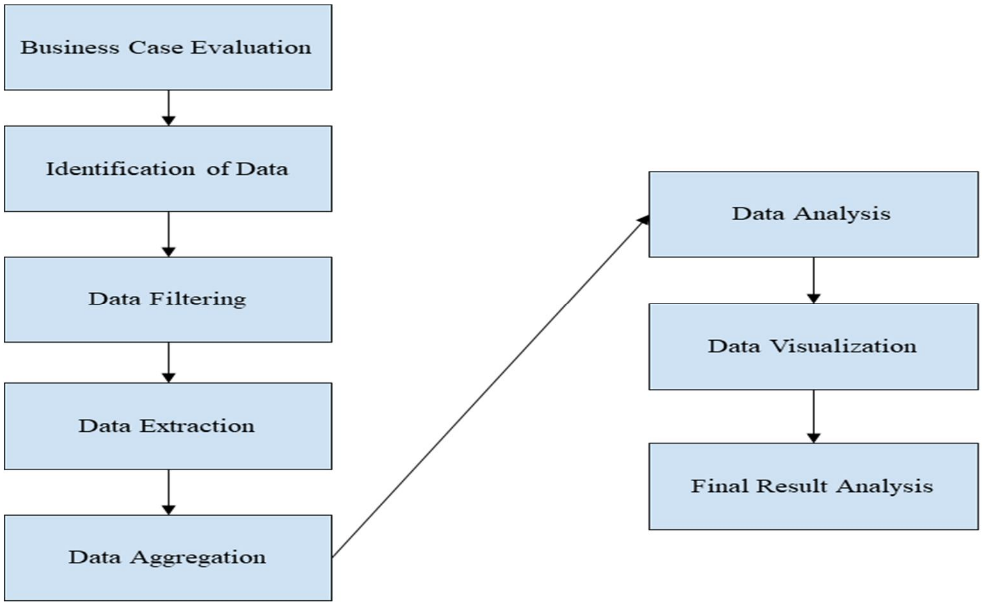

Fig 4. Life cycle of Big Data

Big data analytics lifecycle begins with the business case which defines the reason and the goal behind the analysis. Profound justification and evaluation of big data will advocate the business professionals to get the perception of challenges and resources that will be required to overcome those issues. This will help the team to pry the problems and develop the understanding to formulate the hypothesis.

B. Identification of Data

This is the key step, which broadly specifies the identification of the data source. This step is primarily important because it allows us to get familiar with the data source and helps in getting an understanding, how valuable this data is in order to discern the hidden pattern and correlations.

C. Data Filtering

All the data is accumulated and identified from the previous stage and filtered here to eliminate all the contaminated data. Generally all those data that have no importance in generating the final result and analysis are removed. These include data with high numbers of null values, missing values, columns with zero or negative correlations. These step is carried multiple time, in irregular order, to produce extremely clean dataset

D. Data Extraction

Data extraction step finds out the data that is not congruent with the tools and transforms to the form that is non-conflicting and compatible with big data solutions which can be further used for data analysis.

E. Data Aggregation

This step involves data with the same field across different datasets to be integrated. Although, If the data is voluminous, it can be a bit time consuming due to different data models, but can be executed without any human intervention.

F. Data Analysis

After performing all the prior steps, next comes another chief process to perform exploratory data analysis(EDA). EDA refers to the process of analyzing and probing the dataset to summarize the patterns, to detect irregularities, often by employing graphical representation and summary of the statistics in order to understand the data. By cleaning, modelling, and transforming the data, useful decisions for business solutions are perceived and also, hidden insights from the data, about the business is found.

G. Visualization of Data

Data visualization refers to the technique of graphically summing up the data for communicating the analysis result. Visualization often helps in spotting trends in the dataset. When the data is simply stored in any database or in software, it becomes difficult to make any judgement, but after visualizing tools such as Tableau and Power-BI, these data can be transformed into important graphs and visuals and that is why, data visualization holds the importance

H. Final Analysis Result

In this step, all the decisions are made for the business by the stakeholders through dashboards which allows them to take appropriate actions. By using appropriate machine learning algorithms, decision making processes could be made more precise by automating the learning procedure for the future.

VII. BIG DATA CASE STUDY

Growth and innovation in technology has led companies to leverage its use to make their business thrive fiscally, to get a better understanding of their customer base and find a way to stay ahead in the market. These companies make proper use of Artificial Intelligence, Machine Learning, Cloud computing and Big Data mainly to improve their business solution and generate more revenue. Technology doesn't discriminate among the type of business and endorse all sorts of companies to gain it’s advantage to burgeon. In these case studies, we are going to discuss Starbucks and google.

A. Starbucks

Starbucks is the American multinational chain of coffeehouses which has been rated as one top-notch coffee company in the world with 32,938 store outlets among 64 countries throughout. Profit generated by Starbucks is in billions and the secret ingredient for it’s business triumph is data. Data analytics and Big Data is key for Starbucks success to collect insights and make a profound analysis to make germane pricing strategies, store outlet planning, product development and marketing strategies.[10]

But schemes used by Starbucks have not always been the same. Starbucks has started employing Big Data within recent years, prior to that, like many organizations, decisions made by Starbucks were human based, relied upon experience and perception. The multi billionaire company has faced quite a challenge to establish their empire. Below discussed are key challenges faced by Starbucks-

- Amidst the rapid growing world, Starbucks is facing challenges to deal with its competitors. With numerous renowned coffee chains running in the world namely- costa coffee, Tim Hortons, Mc café etc. It becomes necessary to stand out to be the best to grow the large customer base. Though Starbucks is opening its new outlets in different locations, it’s important to know their rivals and their intentions.
- Monitory is another factor that is affecting the growth of Starbucks. Starbucks coffee are generally more costly than other coffee chains in the word as they use best quality coffee beans for their product and, thus providing it’s opponents with pecuniary advantage.
- Location preferences is the major problem Starbucks is dealing with. In order to open more outlets, it becomes necessary to know whether the location is suitable for alluring a higher number of public. Population of the location is not only the matter, but also if they can afford such extravagant coffee.
- Recently Starbucks provided a new update of mobile order and pay which became quite successful to remotely order the product and pick and go, but has been causing difficulty in service. This has led to long lines and disappointing the customers, who sometimes leave without ordering.
- In year 2021, approximately 1404 stores has been added in USA but due to failure, Starbucks has announced that it will be closing more than 100 outlets in USA and 100 stores in Canada, therefore more number of store is creating major problem to run them efficiently.[11]

1) *Big Data Analytics in Starbucks:* Starbucks always tracks down it’s customers and knows what kind of coffee their buyer prefers, at what point of time and at which location. Starbucks pairs down an individual's data with it’s millions of customers and creates real time, actionable data. Currently, more than 90 millions transactions are happening per week within its 32938 outlets worldwide, where Artificial Intelligence and big data is endorsing with business decisions, sales and direct marketing This section discusses the major area, which helped Starbucks to grow as an individual company-[12]

a) *Reality Decision*: In the calendar year 2008, Starbucks had to shut 100’s of its store locations worldwide. As a solution, to select an apt location Starbucks took a more analytical approach. Today, Starbucks uses Big data to decide if the particular location would be pertinent for the new outlet or not. The analysis is done based on factors including but not limited to-population demographics, accessibility to the location, competition in the vicinity, economic factor, parking adequacy and so on. Starbucks gathers all this data from atlas, which is the BI platform originally originated by Esri(Environmental system research institute), a GIS(Geographic information system) software company. From all the data accumulated, Starbucks evaluates revenue generation from the store and, therefore, decides if the selected location is a feasible option or not.

b) *Personal Offers*: Offers and deals have a huge impact in provoking the customers. Starbucks collects data of an individual and to get an understanding of their preferences while ordering and tracks down the buying pattern in order to send personal offers to each individual. Also, it recommends new products based upon their order, which the customer might enjoy using big data and business intelligence. These allure a huge customer base and create more profitable business.

c) *Optimizing Digital Menus*: Starbucks uses data collected in store as well as research about the customer in the market to adjust and flourish it’s menu with new digital menu boards. Using the previous existing data, the board can make recommendations based on time(breakfast time, lunch time, dinner time etc.), weather(the system analyzes the weather condition and offers beverages accordingly), festivals(the system determines which is the most preferred drink for a particular festival) etc. to boost its sales. Moreover, the board also holds the capacity to effectively make an amendment in the pricing of the product varying from day till night. Thus, digital boards have led Starbucks to reach out to a large customer base.

B. Google

Google is a multinational tech company, which took the charge to collect and organize the information throughout the world and made it publicly accessible and useful. Today, Google processes 20 petabytes of data per day, including 3.5 billion search queries. But these statistics have not always been the same. In its initial years, according to Google, when it was newly introduced to the world, the search queries that were made were just 10,000 per day i.e. 0.0002% of what is being generated today. From then on, the users on the internet has been perpetually increasing, leading to growth of google. In early 2000, Google used a traditional data server to store highly relational data and SQL(standard query language) became the key language of the data. This was considered as one efficient way, which answered all the questions and provided prominent query optimization. But after 2000’s, as the number of internet users rose, the number of queries generated started to discern exponential increase which also amended the form of data.

As workload started to grow, it became a catastrophe for single computers and traditional databases to bear the load and even the most extravagant hardware saw complete fiasco and, therefore demanded a shift from single database nodes to clusters of databases, working concurrently.

1) *Challenges faced by Google*

a) As thousand's of search queries were raised per second, traditional databases were not capable of storing and analyzing the marvelous unstructured data.

b) Every query reads about 100's of MB of data and consumes 10's of billions of CPU cycles and therefore, a single server could not handle this huge data.

c) As thousands of queries were searched per second, it required a large system(to store, process and capture huge amount of data), distributed system(because these data could not rely on just one server even with multiple disks stacked up as machine failure could result in loss of data), highly fault tolerant system(the data or file system runs on 100's or even 1000's of storage machine) to store and process the queries.

2) *Big Data Analytics in Google:*[13]

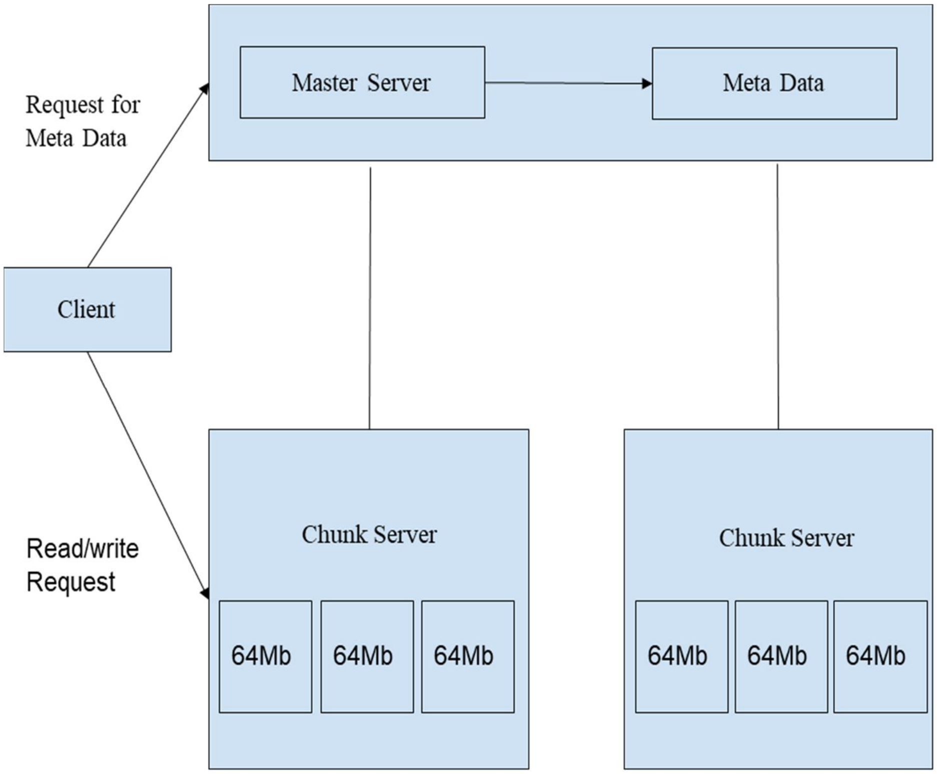

Fig 5. Working of Google File System

As Google faced these issues and setbacks, it became necessary to generate an efficient solution, which can handle all the mentioned challenges. The result was GFS(Google File System). To facilitate GFS, which could store huge amounts of data, there was an architecture developed with one master machine and multiple chunk server or slave machines. Master machine contains metadata i.e. data about data(information about data), while slave machine is responsible for storing data in distributed fashion(replica of data). Any client or an application which would want to read the data will first contact the master slave by placing the request of reading the data(requesting for metadata). Master server, having the data either in RAM or disk Know's in which slave machine the data is stored in distributed fashion. Master machine would respond back with the metadata information to the client and the client could use the information to read/write data to these slave machines, where the actual data is stored. This is the working of GFS.

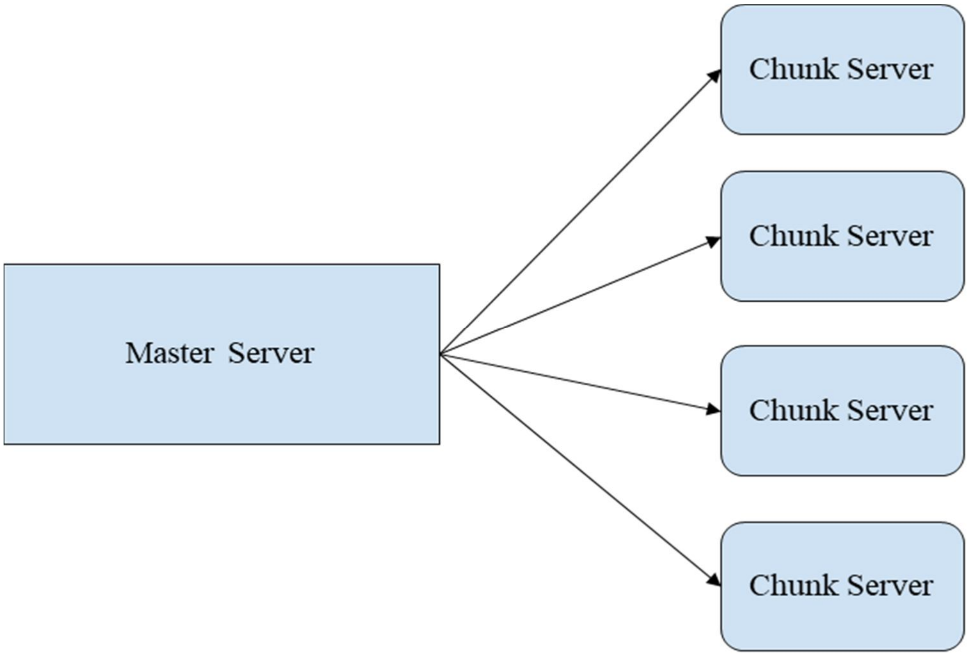

Fig 6. Master-Chunk Architecture

Chunk servers means files getting divided into fixed size, into multiple servers(at most 3 servers), therefore when the amount of data is huge, the file gets divided into multiple chunks or blocks, each block has a fixed size(pre-defined size of 64 mb).

3) *Benefits of GFS:* Benefits of the google file system has provided great support to deal with big data. Below mentioned are key benefits of GFS for storing and managing big data-

- The Master machine has all the information about where the data is stored, location of all the data in different blocks/chunks, and size of the data. It would also get an acknowledgement, when a particular chunk has crashed and where it's replica has been stored.
- The particular file breaks down into different chunks, each chunk contains some part of data which we call as replication of data. This implies, if a particular machine or system crashes, the data will still be present in other chunk servers i.e. no data loss. This whole process helped Google to store and process huge amounts of data in a distributed manner and provided fault tolerant, distributed, scalable storage which could allow them to store huge amounts of data.
- Due to the functionality of horizontal scalability, the system keeps on adding the new nodes as the amount of data keeps on increasing, therefore there is no need to create a larger cluster.

VIII. CONCLUSION

In the research above, we have examined how the methods of storing the big data have transformed over the years and gained lots of attention due to the benefits and opportunities it provides. Disparate challenges faced by traditional databases are profoundly analyzed and the need of NoSQL and Hadoop is thoroughly discussed which has been endorsing different organizations to overcome their business limitations. With effective decision making, big data has grasped complete dominance over the business-technical companies and acted as a spark to drive them further, helping them to deal with their rivals and current market trends. Big data works in complete accordance with data analytics, Machine Learning to make decision making more in-depth to find trends within the data. Different steps taken in order to work with data and making visualization and prediction are discussed rigorously. Additionally, different types of big data analytics are taken into account with real world examples. Lastly, research is carried out on esteemed companies like Starbucks and Google to understand real world use of big data analytics in business and tech firms.

With globalization, in the foreseeable future, the amount of data accumulated will be beyond our imagination. Although Big data technologies are prevailing in use, seeing the current trend of velocity of data gathered, this technology may soon start to discern issues with memory and speed, but again it's uncanny. The world must prepare itself for a more advanced future teemed with innovations, governed by data.

REFERENCES

[1] Simplilearn. “What is Data: Types of Data, and How To Analyze Data?” https://www.simplilearn.com/what-is-data-article. Accessed https://www.simplilearn.com/what-is-data-article.

[2] sas.com. “Big Data What it is and why it matters.” https://www.sas.com/en_in/insights/big-data/what-is-big-data.html.

[3] Elgendy, Nada, and Ahmed Elragal. “Big Data Analytics: A Literature Review Paper.” vol. 8557, 2014, https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper/citation/download.

[4] Internet growth statistics.”History and Growth of the internet from 1995till today.” https://www.internetworldstats.com/emarketing.htm.

[5] Ahmadi, Mohammad, and Parthasarati Dileepan. “A SWOT analysis of big data.” https://www.researchgate.net/publication/303363742_A_SWOT_analysis_of_big_data. Accessed 2016.

[6] Quora.com. “What is the difference between Hadoop and NoSQL?” https://www.quora.com/What-is-the-difference-between-Hadoop-and-NoSQL.

[7] captechu.com. “The Five Key Types of Big Data Analytics Every Business Analyst Should Know.” 2018, https://www.captechu.edu/blog/five-types-of-big-data-business-analytics.

[8] KDnuggets. “Introduction to Blockchains & What It Means to Big Data.” https://www.kdnuggets.com/2017/09/introduction-blockchain-big-data.html.

[9] TutorialsPoint. “Big Data Analytics - Data Life Cycle.” https://www.tutorialspoint.com/big_data_analytics/big_data_analytics_lifecycle.htm.

[10] James, Lindsay. “Big Data: The Secret to Starbucks’ Supply Chain Success.” 2020, https://www.sisense.com/blog/big-data-the-secret-to-starbucks-supply-chain-success-2/.

[11] UKESSAYS. “Critical Analysis of Strategic Issues faced by Starbucks.” 2021, https://www.ukessays.com/essays/marketing/critical-analysis-of-the-strategic-issues-faced-starbucks-marketing-essay.php.

[12] Marr, Bernard. “Starbucks: Using Big Data, Analytics And Artificial Intelligence To Boost Performance.” 2021, https://bernardmarr.com/starbucks-using-big-data-analytics-and-artificial-intelligence-to-boost-performance/.

[13] Ghemawat, Sanjay.”The google file system.” https://storage.googleapis.com/pub-tools-public-publication-data/pdf/035fc972c796d33122033a0614bc94cff1527999.pdf.