Cloud Based Big Data Analytics

Devanshi Goel
Institute of Technology
Nirma University
Ahmedabad, India
19bce067@nirmauni.ac.in
Ph. no: 9376706563

Mahek Parekh
Institute of Technology
Nirma University
Ahmedabad, India
19bce154@nirmauni.ac.in
Ph. no: 9879787801

ABSTRACT:

Every day, an enormous quantity of data is generated from many sources. Because this data is so dynamic and large in size, conventional processing techniques are unable to handle it. Big data is the storage, processing, and analysis of massive quantities of data. Large scale and complex computing can be performed very well by cloud computing. We can eliminate the requirement to maintain high-cost computing H/W, storage, and software with the aid of cloud computing technologies. Hence, cloud computing can offer infrastructure to perform Big Data Analytics in a cost effective and efficient manner. The current state of research, problems, concerns, and future research directions in this topic of Cloud Based Big Data Analytics are presented in this paper.

Keywords: Cloud Based Big Data Analytics, Cloud Computing, Big Data Analytics, Data Analytics as a Service, DAaaS

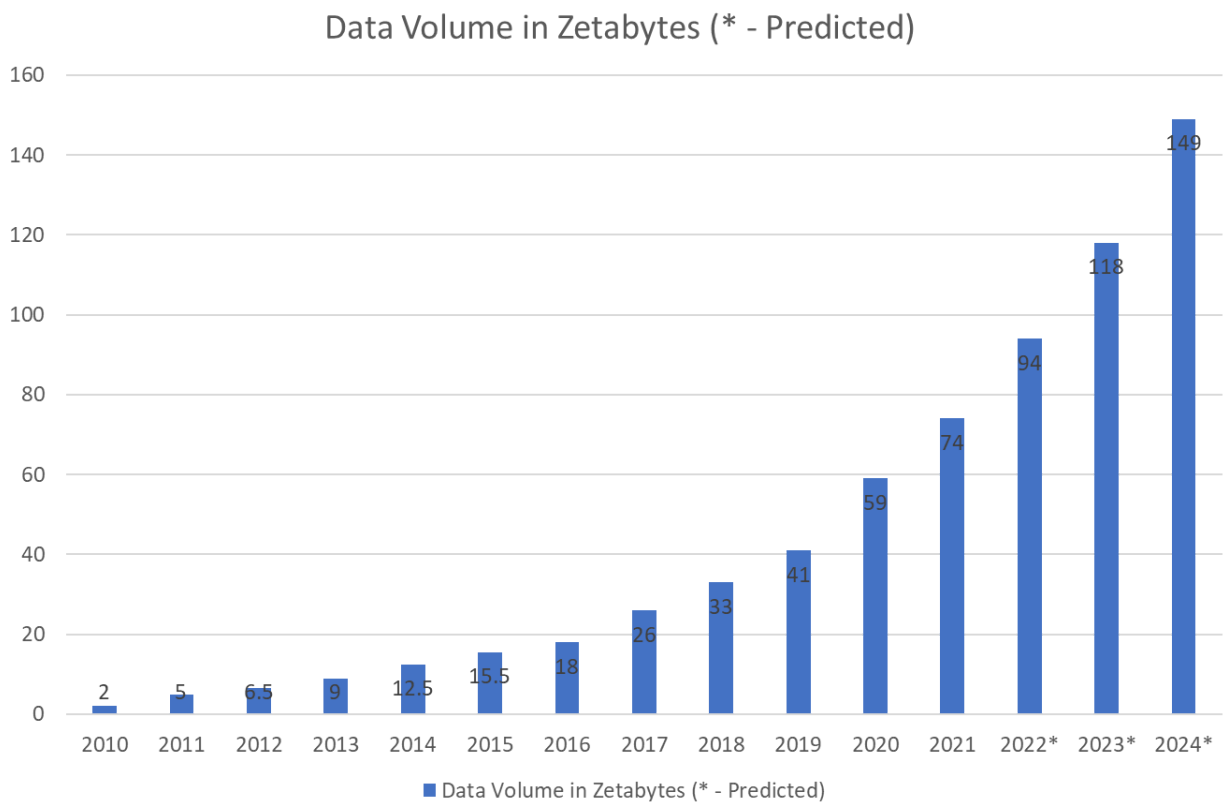
INTRODUCTION:

SR	Authors Name	Year of Publication	Streaming Data	Big Data Analytics	Distributed Computing	Storage and warehouse	Hadoop
1	Subia Saif, Samar Wazir	2018	Yes	No	No	Yes	Yes
2	Samiya Khan, Kashish Ara Shakil and Mansaf Alam	2015	No	Yes	No	No	No
3	Ibrahim Abaker	2014	Yes	No	Yes	No	No
4	R.Suganya, M. Pavithra, A.Rathika, R.Ashwini	2019	No	No	No	Yes	No
5	Dr. Narasimha Rao Vajjhala, Dr. Ervin Ramollari	2016	No	No	No	No	Yes
6	Amitkumar Manekar, Dr. G. Pradeepini	2015	Yes	No	Yes	No	No
7	Affreen Ara, Aftab Ara	2016	No	Yes	No	Yes	No
8	Manoj Muniswamaia, Tilak Agerwala and Charles Tappert	2019	No	No	Yes	No	Yes

We live in the information age. We see data everywhere because of wonderful technical advancements that have taken place in recent years. Data can be collected from a variety of sources such as smartphones, social networks, Internet of Things(IoT), sensor devices, servers, etc. Hence, the global amount of data is increasing exponentially because of the increase in these devices that collect the data.

Some statistics:

- 20-30 years ago only 1% information produced was digital whereas today it is 94%.
- Globally 44 zettabytes(44*10^21 bytes) of data generated daily and it is expected to increase exponentially to 163 zettabytes(163*10^21 bytes) till the year 2025.
- Google generates more than 5,00,000 searches every second
- Every minute
    - 5,27,760 photos shared in Snapchat
    - 120 people join LinkedIn
    - 4,14,66,000 YouTube videos get watched
    - 46,740 photos are posted on Instagram



With the advantages, we also face many challenges with the increase in quantity of data as conventional systems like relational databases struggle in processing and analyzing data of huge size. As a result, the phrase "Big Data" was coined to characterize both the volume of data and the necessity for new technology to manage and analyze it.

WHAT IS BIG DATA - DEFINITION AND CHARACTERISTICS:

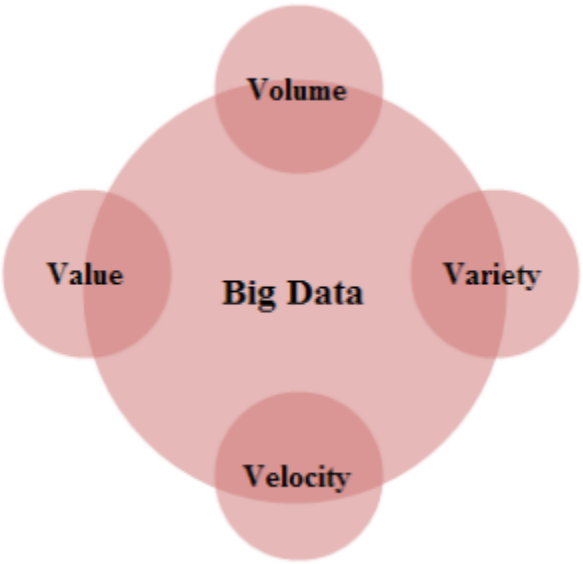
Big data means a very large volume of data. Generally the data which are categorized as big data are not at all easy to store, analyze or process by means of normal traditional databases. It is because of the fact that the data is so large in size such that the normal database query would take so long to respond to the request that the company would lose its customer base if it will depend on that.

Big Data is considered as too big(data coming from various sources at petabyte level), too fast(data growth is fast and must be processed quickly) and too hard(data not adapting to conventional tools) for existing technologies to process.

Three characteristics of big data are:

1) Data is very large
2) Regular Relational Database cannot categorize the data
3) Data is very rapidly produced, captured and processed

Many sectors such as healthcare, business, science, finance, engineering, etc. are being transformed positively with the help of big data. Hence, eventually society too get transformed.



Four V's which describes and characterize big data:

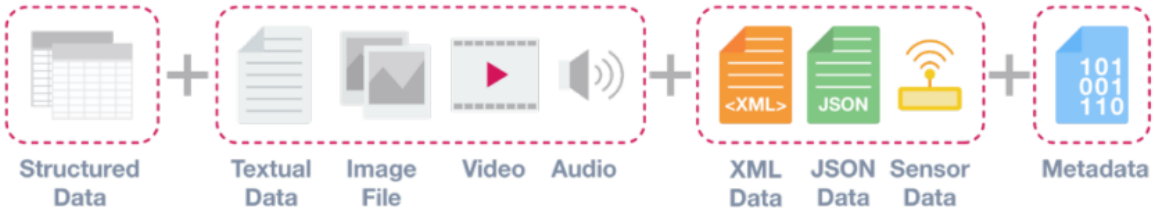
1. **Volume:**

It refers to the entire amount of data generated from a variety of sources, which is always increasing. One of the benefits of gathering large amounts of data is the ability to uncover hidden information and patterns through data analysis. Longitudinal data collection takes a lot of effort and requires a lot of money.

2. **Variety:**

Variety speaks for itself. It means the different devices or machines through which we can collect data. Sensors, cellphones, and social media networks are all examples of this. Such data types include video, image, text, audio, and data logs in both structured and unstructured formats.

Unstructured data makes up the majority of the data created by mobile apps. Text messages, online games, blogs, and social media, for example, provide a lot of unstructured data thanks to mobile devices and sensors. Internet users also produce a vast array of organized and unstructured data.



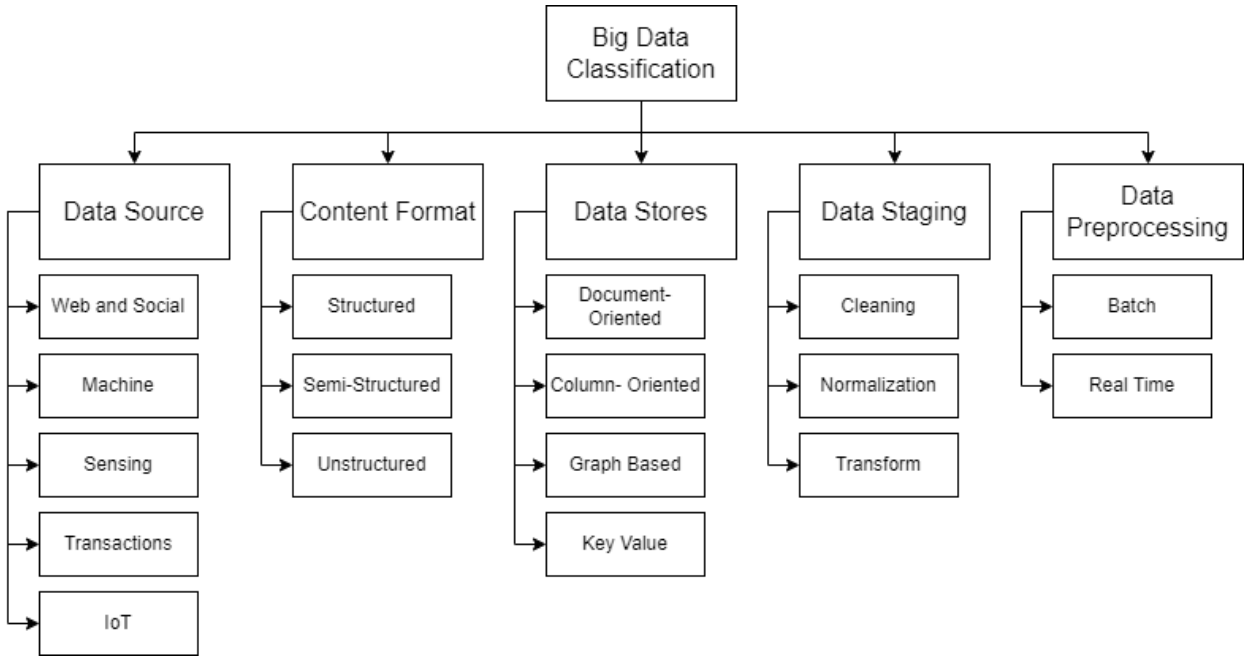
3. **Velocity:**

Velocity basically means how fast the data is transferred. As data is dynamic and it changes every 100th of a second, we need to update it accordingly. Because of the absorption of supplementary data collections, streaming data arriving from numerous sources, the contents of data are continually changing.

4. **Value:**

It is by far the most important aspect of big data. It refers to a technique for swiftly extracting massive hidden values from large datasets of various types.

CLASSIFICATION OF BIG DATA:



Classification based on Data Sources:

- Social Media: It is data produced via Uniform Resource Locator(URL) to exchange ideas.
- Machine Generated Data: information produced automatically from devices like computers without interference of humans.
- Sensing: devices that monitor and transfer physical values into signals
- Transactions: data which includes timestamp for describing data
- IoT: devices that are individually recognized as a part of the internet like smart door locks, etc.

Classification based on Content Format:

- Structured: Generally it is managed SQL which input, query, store, and analyse data with ease. Eg. words, numbers and dates.
- Semi Structured: Some data that not follows traditional DB or database systems like relational database models. Eg. Table, etc.
- Unstructured: Data that do not follow any specified format like text messages, location information, videos, social media data, etc. As the use of smartphones is increasing, the size of this sort of data is continuously increasing.

Classification based on Data Stores:

- Document Oriented: mainly to store and recover collection of documents
- Column Oriented: stores its contents in columns apart from rows
- Graph Database: stores and represents a data that uses a graph model with vertices and edges
- Key Value: alternative to relational database system where cells are defined by a row and a column key

Classification based on Data Staging:

- Cleaning: process which helps in finding missing and unreasonable data
- Transform: process which helps in altering data into a format that may be used for analysis
- Normalization: process of organizing database to reduce duplication

Classification based on Data Processing:

- Batch: data is processed in batch of jobs
- Real Time: data is processed in real time

HOW IS CLOUD COMPUTING RELATED TO BIG DATA:

Data Analytics As A Service (DAaaS):

For storing and processing huge volumes of data, scalability, fault tolerance, and availability are all essential. Cloud computing, through hardware virtualization, may supply all of these. As a result, big data and cloud computing are mutually beneficial concepts, because the cloud makes massive data accessible, scalable, and fault-tolerant.

Businesses regard big data as a possible economic opportunity. As a result, new companies have sprung up to provide Big Data as a Service (BDaaS) or DataBase as a Service (DBaaS) solutions. Big data is also available on demand from businesses like Google, IBM, Amazon, and Microsoft.

Cloud computing is a trend in technological growth, as technological advancement has accelerated the development of the electronic information society. The phenomenon of big data emerges as a result of this, and the rapid increase of big data is a concern that may occur as the electronic information society evolves. Because big data is concerned with storage capacity in the cloud system, and cloud computing makes use of large computational and storage capacities, the two go hand in hand. As a result, by offering processing capabilities to big data applications, big data stimulates and accelerates the development of cloud computing.

Big Data Analytics is a method of analyzing large amounts of data in order to find hidden patterns, unknown relationships, market trends, client preferences, and other important information. Data Analytics as a Service (DAaaS) is an extensible analytical platform delivered via a cloud-based delivery mechanism, with numerous data analytics tools available and configurable by the customer to process and analyze large amounts of heterogeneous data quickly. The combination of Big Data and Cloud can help firms embrace analytic capabilities across vast volumes of heterogeneous data sources that they can't handle on their own.

The value gained from big data analytics is mostly beneficial to businesses. The value of the data lies in the data itself, not in the skill that went into it. For the future, businesses must have access to services that provide advanced big data analytics. They must also establish solid infrastructure, as well as the necessary capital investment. Because of its elasticity, the cloud may be a valuable resource for analytics engineers. Analytics service providers have modified the Software as a Service (SaaS) model to assist cloud in solving big data concerns. The user does not need to buy all of the gear in the Software as a Service approach; all they need is a web browser. Because consumers need to install extra software resources such as Hadoop and Cassandra and other basic components to achieve greater performance from analytics companies, big data platforms are complicated platforms.

DAaaS Applications:

Data analytics as a service is a business that is altering trends since it has a wide range of applications. By adopting and utilizing DAaaS, businesses get a competitive advantage. Analytical approaches in various fields may differ. The following are the DAaaS applications:

Social Media:

Cloud data analytics is extremely important in social media and websites like Facebook and Twitter. Analytics are used on these websites to analyze user social media activity and consumer behavior. Cloud drivers enable simultaneous analysis of social media data, with the results analyzed in real time.

Medical:

Big data, cloud, and analytics technology can assist doctors in making better judgments by allowing them to do real-time analysis of patient history, genome sequencing, blood tests, and research similar medical cases and research articles from around the world. Doctors can reduce treatment costs by using analytic tools, strategies, and maximizing the use of technology.

Smart Cities:

Smart cities may benefit from big data and can use cloud-based analytic services to assist generate information intelligence. Transport, energy, smart lighting, noise pollution, citizen information, trash management, water management, and other real-time data may all be supplied into the cloud for analysis. To offer better governance, planning, and administration, city management can obtain information from data acquired from numerous sources. Cloud-based big data analytic solutions may enhance and benefit areas of urban life such as health, pollution, public transportation, and air pollution.

Keeping track of product preferences and sales:

To follow items across its warehouse, Amazon leverages data analytics on cloud drivers. Users can get suggestions from websites based on their prior search history. The majority of user preferences are saved in the cloud for future usage. Data from throughout the network is analyzed by social media analysts to give a benchmark for markets to measure their own performance.

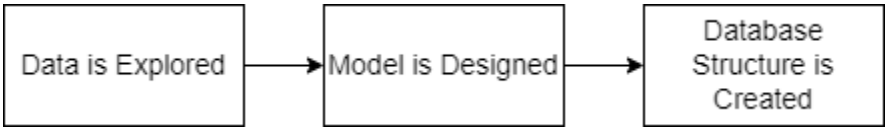
Maintaining customer records:

websites save user profiles on the cloud for future use, and they capture and analyze data in real time on local servers. The majority of user data is saved on the cloud, so that when a user switches computers, his preferences do not alter.

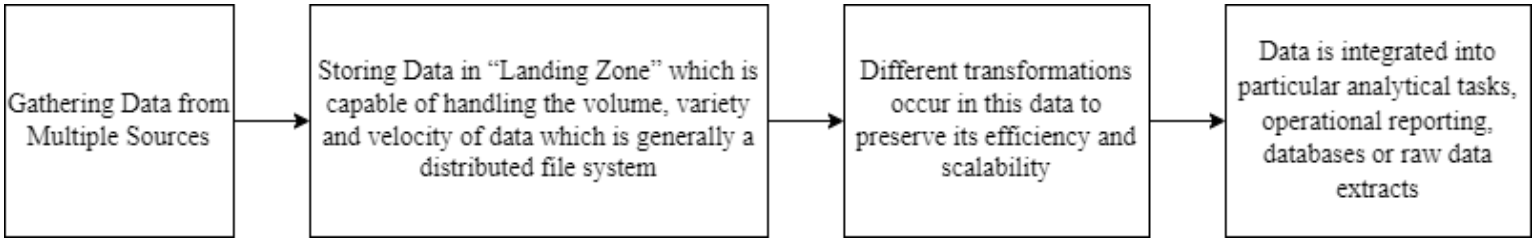
Big Data Analytics Cycle:

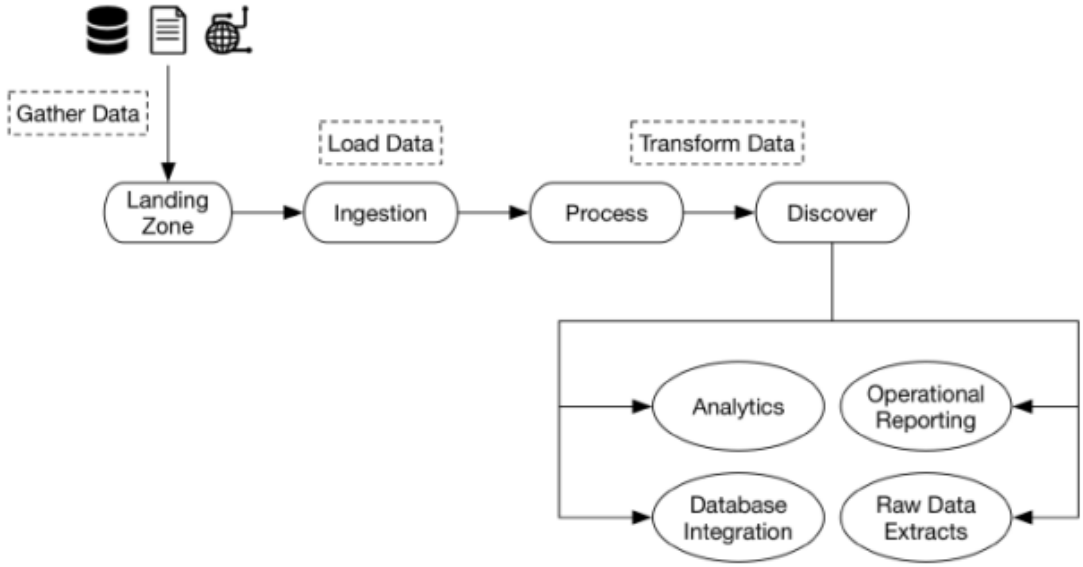
The processing step in big data analytics is different from the processing step for conventional transactional data.

In conventional environment, the steps are as follows:



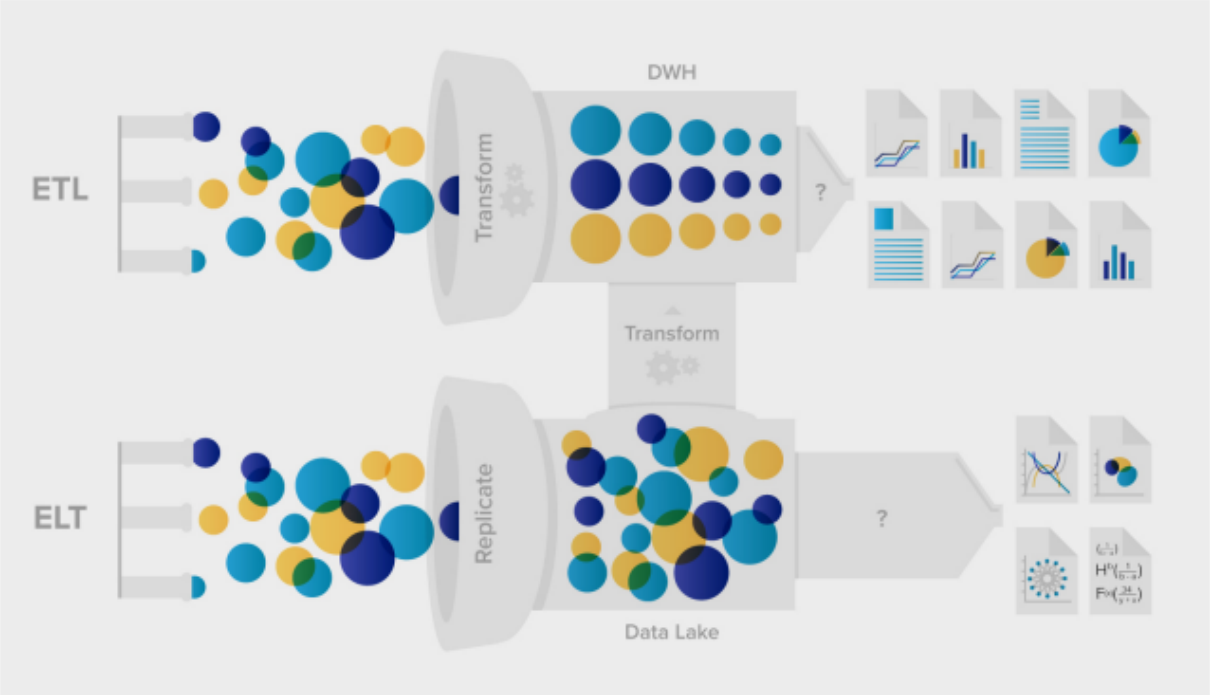
In Big Data Analysis, the steps are as follows:



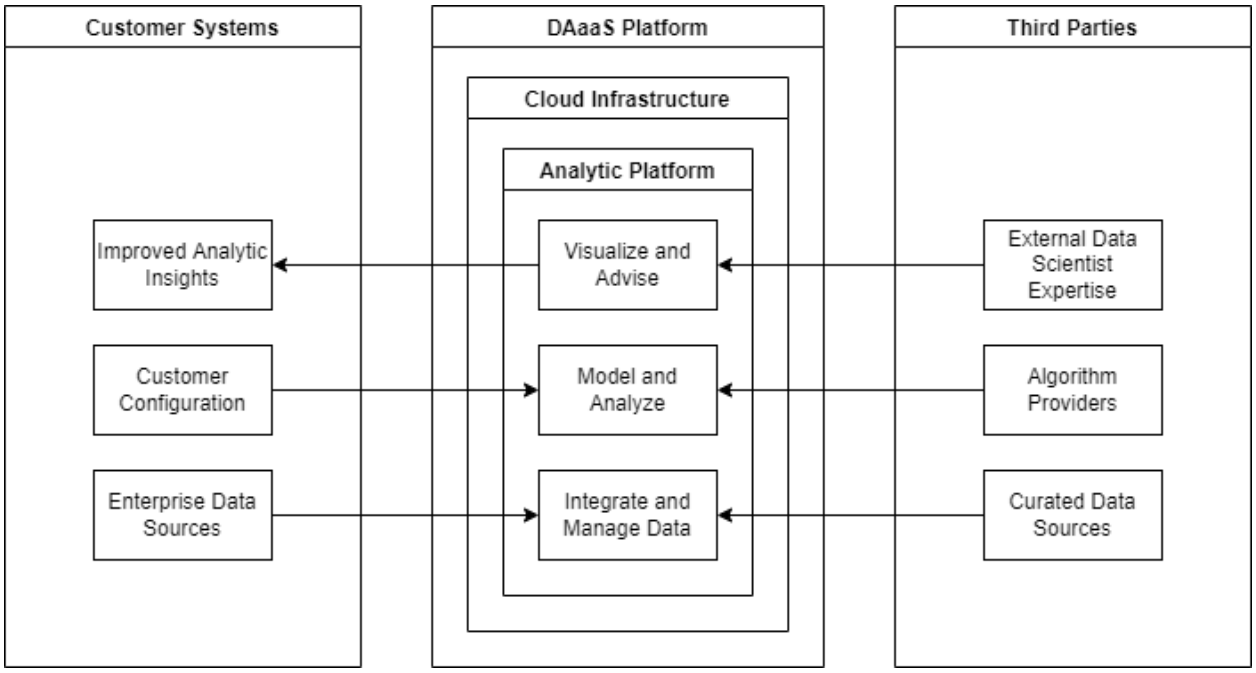


Moving from ETL to ELT Paradigm:

- ETL (Extract -> Transform -> Load): This strategy has the disadvantage of requiring a lot of Input Output activity, string processing, variable transformation, and data parsing.
- ELT (Extract -> Load -> Transform): ELT offers advantages over ETL because it takes the most computationally expensive job that is transformation, and moves it to the cloud rather than an on-premise service that is already overburdened with routine transaction processing. As a result, we apply the ELT technique in Cloud Based Big Data Analytics.



Architecture Diagram of Cloud Based Big Data Analytics:



This combination can benefit the entire world and provide a significant analytics edge for generating information that is perfect for business continuity. Let's take a look at the benefits of merging big data and cloud computing for businesses.

1. **Agility:**

Because storing and managing data is a time-consuming and arduous operation, traditional systems have proven to be slower. Since the introduction of cloud by businesses, it has provided all of the resources needed to run many virtual servers in a cloud database in real time.

2. **Affordability:**

A private data center is very costly as businesses have to pay for the hardware, power, ongoing maintenance, etc. The cloud can provide all such services on request on rent as per pay as you go model.

3. **Accessibility:**

Cloud resources and services are present in major regions of the world. Hence, data and processing work can take place nearer to the region where the big data task is located. For example, if a big data task is situated in some region, it is relatively easy to implement the resources and services for the project in that specific region instead of bearing the cost of moving the data to another region.

4. **Resilience:**

Cloud Resiliency means capability of cloud resources such as servers, storage system, network, etc. to recover without any data loss and continue their operations without systems connected to the network going down.

Data is most important in any big data project. Benefit of cloud resilience in cloud based big data analysis is data storage reliability. This can be achieved as cloud replicates the data to maintain durable storage.

Now let's see the cons or major disadvantages of merging big data and cloud computing.

1. **Network Dependence:**

For service providers, big data analysis on the cloud poses networking challenges. To bring data to an analytics team in a timely manner, technical competence is required.

2. **Storage:**

Virtualization loads influence storage, which has a negative impact on performance in a heavily virtualized and dispersed cloud.

3. **Security:**

Businesses may be worried that their data is being held on a virtual server and being passed on to a third party. Because data becomes public after analytics, privacy can have an influence on cloud analytics.

4. **Lack of Standardization:**

The tools used to collect large data sets might be inaccurate at times.

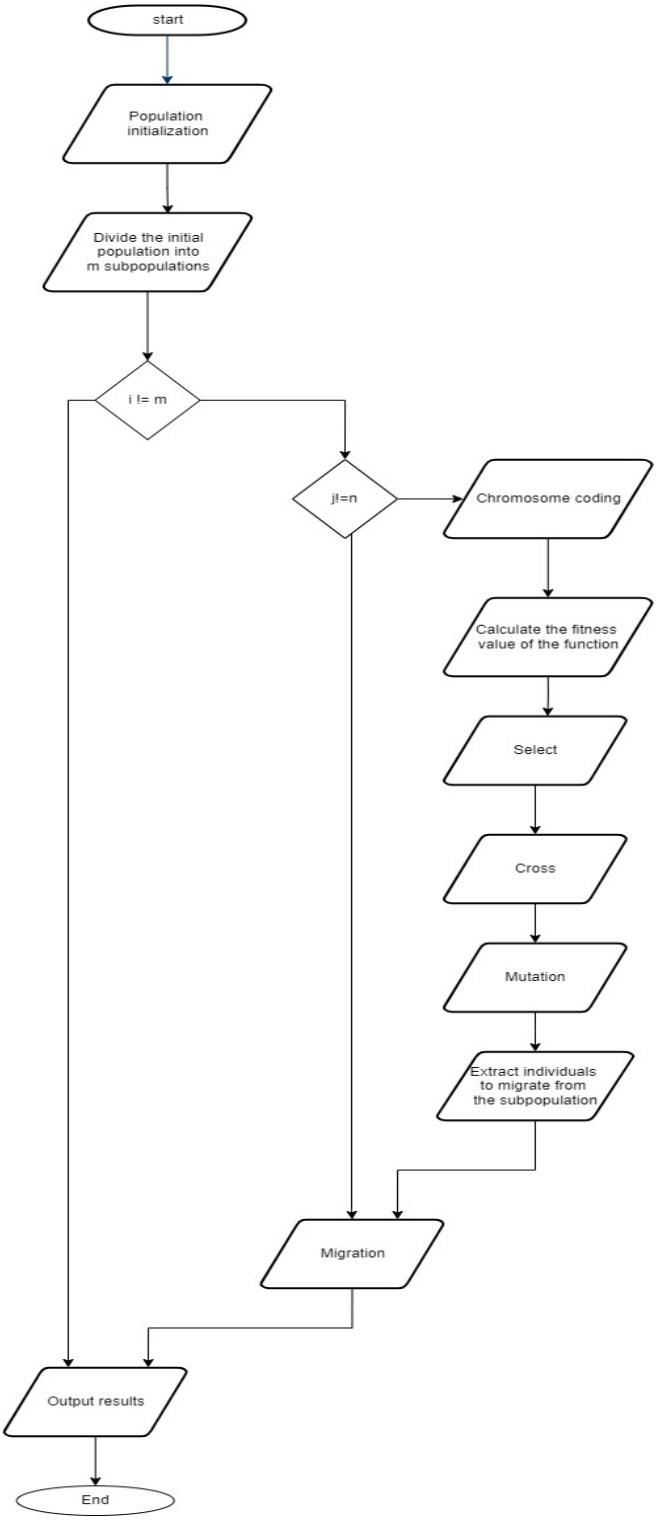
5. **Analytic Knowledge:**

Analytics need a skilled data scientist who is up to speed on the most cutting-edge analytics methodologies.

6. **Data Volume:**

Moving data to the cloud for computation becomes a strenuous task but vice versa is not true.

ALGORITHM:


start
Population initialization
Divide the initial population into m subpopulations
i != m
j!=n
Chromosome coding
Calculate the fitness value of the function
Select
Cross
Mutation
Extract individuals to migrate from the subpopulation
Migration
Output results
End

CONCLUSION:

Companies that use cloud-based big data analytics will make better judgments and earn more money. Many firms are working on cloud-based big data analytics tools. Companies will be able to implement, administer, manage, and protect big data tools and technologies without difficulty. Companies who do not capitalize on big data business potential will be left behind by their competitors. Companies will examine data acquired from mobile devices, storage arrays, and management tools in the best-case scenario. Big data and the cloud Network optimization and intelligence systems can employ analytics to detect possible intrusions before they happen. When the Internet of Things is linked with analytics, it may help diverge data sources and enable analytics to be used in a variety of areas, such as monitoring sensors.

However, it's uncertain if today's businesses are prepared to capitalize on the Internet of Things' promise. IoT will extend data sources and enable analytics to be applied to new domains, such as sensor monitoring on manufacturing devices. By acting as a large source of data, the Internet of Things might be a benefit to analytics. The advent of new tools and technology will allow smaller businesses to outperform larger businesses in terms of efficiency and productivity, resulting in the creation of new start-ups. More positions for data scientists and analysts would be created as the Internet of Things, data analytics, and cloud grow.

REFERENCES:

1. Subia Saif, Samar Wazir. “Performance Analysis of Big Data and Cloud Computing Techniques: A Survey”

2. Samiya Khan, Kashish Ara Shakil and Mansaf Alam. “Cloud Based Big Data Analytics - A Survey of Current Research and Future Directions”

3. Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan. “The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues”

4. R.Suganya, M. Pavithra, A.Rathika, R.Ashwini. “Cloud Based Big Data Analytics: A Survey of Current Research and Future Directions”

5. Dr. Narasimha Rao Vajjhala, Dr. Ervin Ramollari. “Big Data using Cloud Computing - Opportunities for Small and Medium-sized Enterprises”

6. B. Sc. Blend Berisha, B. Sc. Endrit Meziu. “Big Data Analytics in Cloud Computing: An Overview”

7. Amitkumar Manekar, Dr. G. Pradeepini. “A Review on Cloud Based Big Data Analytics”

8. Affreen Ara, Aftab Ara. “Cloud for Big Data Analytics Trends”

9. Manoj Muniswamaiah, Tilak Agerwala and Charles Tappert. “Big Data in Cloud Computing Review and Opportunities”

10. Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu and Fei Hu. “Big Data and Cloud Computing: Innovation, Opportunities and Challenges”