SURYA ENGINEERING COLLEGE,ERODE
Approved by AICTE,New Delhi & Affiliated to Anna University,Chennai)
Erode-Perundurai Highway Mettukadai,Erode-638107.
Ph.0424-2555018 Mobile No:9842511455 Email:secerode@gmail.com

NAAN MUDHALVAN IBM PROJECT

BIGDATA ANALYSIS WITH IBM CLOUD DATABASES

TEAM MEMBERS

M.VIGNESHWARALINGAM(732821104045)
G.MOHANRAJ(732821104026)
U.RAJESH(732821104032)
V.SRIDHAR(732821104037)
B.GANISHKA(732821104012)

Abstract:

The intersection of Big Data and Cloud Computing is a focal point in the IT industry, driven by the relentless generation of vast and diverse data from various sources. This data's sheer volume, velocity, and variety pose challenges for traditional processing tools. Big Data encompasses the storage, processing, and analysis of these massive datasets, while Cloud Computing provides the cost-effective and efficient infrastructure to support these endeavors. Various sectors, including healthcare and education, are harnessing Big Data's potential to reduce costs, predict pandemics, and prevent diseases. This paper provides an in-depth exploration of Big Data Analytics, starting with an introduction to the concept, the scale of daily data generation, and its defining characteristics.

TECHNOLOGIES USED:

- Hadoop
- Apache Spark
- Data Warehouses
- Hive and Pig
- Apache Flink
- HBase
- NoSQL Databases

Hardware and Software used

Hardware:

- Clustered servers
- High performance storage
- Memory(RAM)
- Distributed file system

Software

- ✓ Operating system
- ✓ Cluster management
- ✓ Query and analysis tools
- ✓ Machine learning libraries

INTRODUCTION :

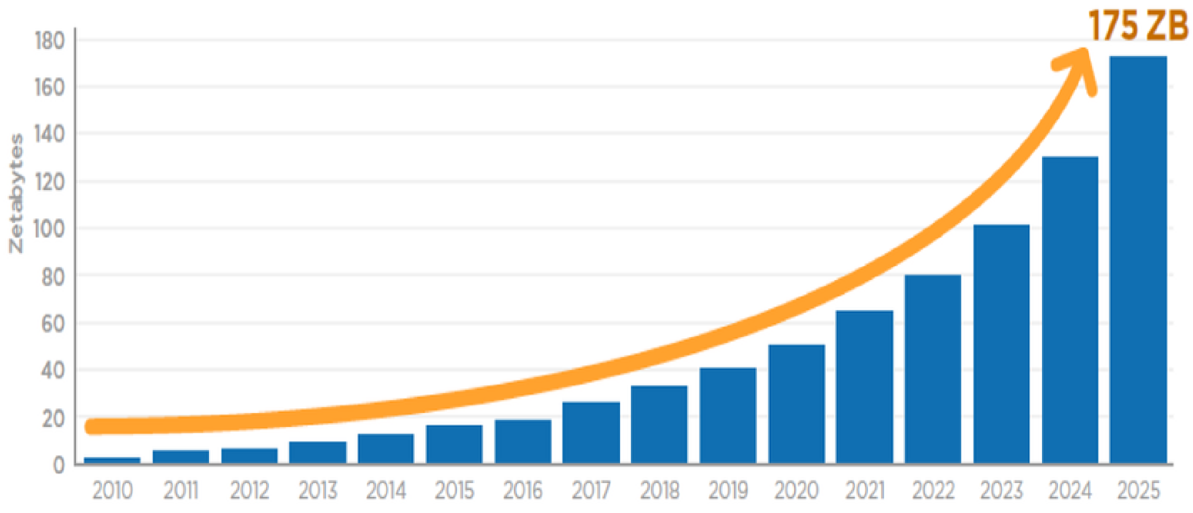
OVERIVEIW:

I. TECHNOLOGIES DEVELOPMENT :

We live in the data age. We see them everywhere and this is thanks to the great technological developments that have taken place in recent years. The rate of digitalization has increased significantly and now we are rightly talking about ”digital information societies”. If 20 or 30 years ago only 1% of the information produced was digital, now over 94% of this information is digital and it comes from various sources such as our mobile phones, servers, sensor devices on the Internet of Things, social networks, etc.

The number and amount of information collected has in- creased significantly due to the increase of devices that collect this information such as mobile devices, cheap and numerous sensor devices on the Internet of Things (IoT), remote sensing, software logs, cameras, microphones, RFID readers, wireless sensor networks, etc.According to statistics, the amoun of data generated / day is about 44 zettabytes (44 × 1021 bytes). Every second, 1.7 MB of data is generated per person Based on International Data Group forecasts, the global amount of data will increase exponentially from 2020 to 2025 with a move from 44 to 163 zettabytes. Amount of global data generated, copied and consumed.

As can be seen, in the years 2010-2015, the rate of increase from year to year has been smaller, while since 2018, this rate has increased significantly thus making the trend exponential in nature.



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

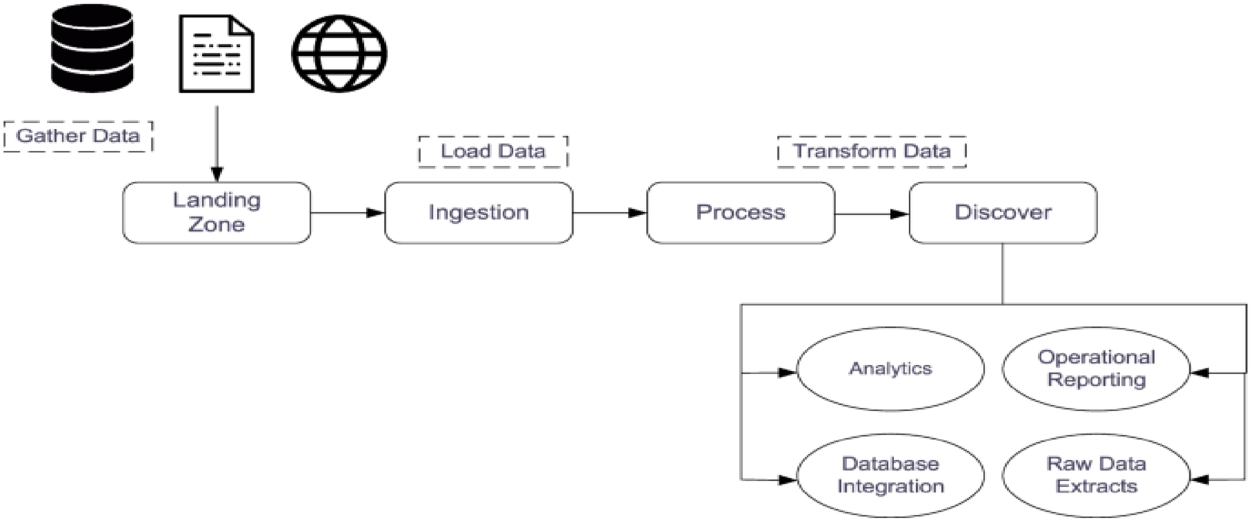
II. BIG DATA ANALYTICS IN CLOUD DATABASES:

Cloud databases can scale horizontally or vertically to accommodate large datasets and high query workloads. Pay-as-you-go pricing models in the cloud allow you to only pay for the resources you use, making it cost-effective for big data analysis.Cloud databases often employ distributed processing techniques to analyze data in parallel, speeding up query performance.You can easily integrate various data sources into cloud databases, enabling comprehensive analysis.Utilize tools like machine learning, data warehousing, and business intelligence services in the cloud for advanced analysis.Cloud providers offer robust security features and compliance certifications to protect sensitive data.

Pair your cloud database with data visualization tools to gain insights from your analysis results.

III. DATA PREPARATION :

Data preparation is the foundational process in any data analysis endeavor. It involves collecting, cleaning, and transforming raw data into a structured and clean dataset that is ready for analysis. During this critical phase, data is carefully examined for missing values, duplicates, and errors, and appropriate actions are taken to rectify them. Categorical variables may be encoded, numerical data scaled, and new features engineered to enhance the dataset's informativeness.



Data formatting ensures consistency and compatibility with analysis tools, while data validation confirms that the data aligns with predefined expectations and rules. Furthermore, data preparation encompasses aspects of security and privacy, ensuring sensitive information is safeguarded, and legal requirements are met. Thorough documentation of each step is essential for transparency and reproducibility. In essence, data preparation is the cornerstone upon which meaningful insights and reliable analysis are built, making it a pivotal stage in the data analysis process.

IV.BIG DATA ANALYTICS CYCLE:

According to processing big data for analytics differs from processing traditional transactional data. In traditional environments, data is first explored then a model design as well as a database structure is created. depicts the flow of big data analysis. As can be seen, it starts by gathering data from multiple sources, such as multiple files, systems, sensors and the Web. This data is then stored in the so called ”landing zone” which is a medium capable of handling the volume, variety and velocity of data.

This is usually a distributed fil system. After data is stored, different transformations occur in this data to preserve its efficiency and scalability. Afer that, they are integrated into particular analytical tasks, operational reporting, databases or raw data extracts .

V . CONCLUSION:

As more and more data is generated and collected, data analysis requires scalable, flexible, and high performing tools to provide insights in a timely fashion. However, organizations are facing a growing big data environment, where new tools emerge and become outdated very quickly. Therefore, it can be very difficult to keep pace and choose the right tools.

MERITS OF BIG DATA ANALYTICS

- ✓ Cost optimization
- ✓ Efficiency improvements
- ✓ Innovation
- ✓ Education .

DEMERITS OF BIG DATA ANALYTICS

- ✓ Privacy and security concerns
- ✓ Technical challenges and requirements
- ✓ A talent gap
- ✓ Security hazard
- ✓ Adherence.