SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS508 - BIG DATA ANALYTICS

III YEAR / V SEMESTER

Unit 1- INTRODUCTION TO BIG DATA

Topic 4 : Perception and Quantification of Value - Understanding Big Data Storage

Perception and Quantification of Value - Understanding Big Data Storage

•Two facets of the appropriateness of big data,

- first being organizational fitness and
- second being suitability of the business challenge.

•The third facet must also be folded into the equation, and that is big data's contribution to the organization.

Cont..

- These facets drill down into the question of value
  - **Increasing revenues**: As an example, an expectation of using a recommendation engine would be to increase same-customer sales by adding more items into the market basket.
  - **Lowering costs:** As an example, using a big data platform built on commodity hardware for ETL would reduce or eliminate the need for more specialized servers used for data staging, thereby reducing the storage footprint and reducing operating costs.
  - **Increasing productivity**: Increasing the speed for the pattern analysis and matching done for fraud analysis helps to identify more instances of suspicious behavior faster, allowing for actions to be taken more quickly and transform the organization from being focused on recovery of funds to proactive prevention of fraud
  - **Reducing risk:** Using a big data platform or collecting many thousands of streams of automated sensor data can provide full visibility into the current state of a power grid, in which unusual events could be rapidly investigated to determine if a risk of an imminent outage can be reduced.

UNDERSTANDING BIG DATA STORAGE

•Most big data applications achieve their performance and scalability through

- **Deployment on a collection of storage and computing resources** bound together within a runtime environment

- The ability to design, develop, and implement a big data application is directly dependent on
  - an awareness of the architecture of the underlying computing platform,
    - both from a hardware and more importantly from a software perspective.

Cont..

- Commonality among the different appliances and frameworks
  - Adaptation of tools to leverage the combination of collections of four key computing resources:
    - Processing capability, often referred to as a CPU, processor, or node.
    - Memory, which holds the data that the processing node is currently working on. Most single node machines have a limit to the amount of memory.
    - Storage, providing persistence of data—the place where datasets are loaded, and from which the data is loaded into memory to be processed.
    - Network, which provides the “pipes” through which datasets are exchanged between different processing and storage nodes

What is Big Data Storage?

- Big Data Storage is a new technology poised to revolutionize how we store data.
- The technology was first developed in the early 2000s when companies were faced with storing massive amounts of data that they could not keep on their servers.
- The problem was that traditional storage methods couldn't handle storing all this data, so companies had to look for new ways to keep it. That's when Big Data Storage came into being.
- It's a way for companies to store large amounts of data without worrying about running out of space.

Big Data Storage Challenges

•The first challenge is how much storage you'll need for your extensive data system.

•If you're going to store large amounts of information about your customers and their behavior, you'll need a lot of space for that data to live.

•It's not uncommon for large companies like Google or Facebook to have petabytes (1 million gigabytes) of storage explicitly dedicated to their big data needs, and that's only one company!

•Another challenge with big data is how quickly it grows. Companies are constantly gathering new types of information about their customer's habits and preferences, and they're looking at ways they can use this information to improve their products or services

Data Storage Methods

- Warehouse Storage

Warehouse storage is one of the more common ways to store large amounts of data, but it has drawbacks.

- Cloud Storage

Cloud storage is an increasingly popular option since it's easier than ever to use this method, thanks to advancements in technology such as Amazon Web Services (AWS)

Data Storage Technologies

- Apache Hadoop,
- Apache HBase,
- Snowflake are three big data storage technologies often used in the data lake analytics paradigm

Data Storage Technologies

- Apache Hadoop,
- Hadoop has gained considerable attention as it is one of the most common frameworks to support big data analytics.
- Apache Hbase
- HBase, you can use a NoSQL database or complement Hadoop with a column-oriented store. This database is designed to efficiently manage large tables with billions of rows and millions of columns.
- The performance can be tuned by adjusting memory usage, the number of servers, block size, and other settings.

- Snowflake
- Snowflake for Data Lake Analytics is an enterprise-grade cloud platform for advanced analytics applications built on top of Apache Hadoop.
- It offers real-time access to historical and streaming data from any source and format at any scale without requiring changes to existing applications or workflows.

Activity

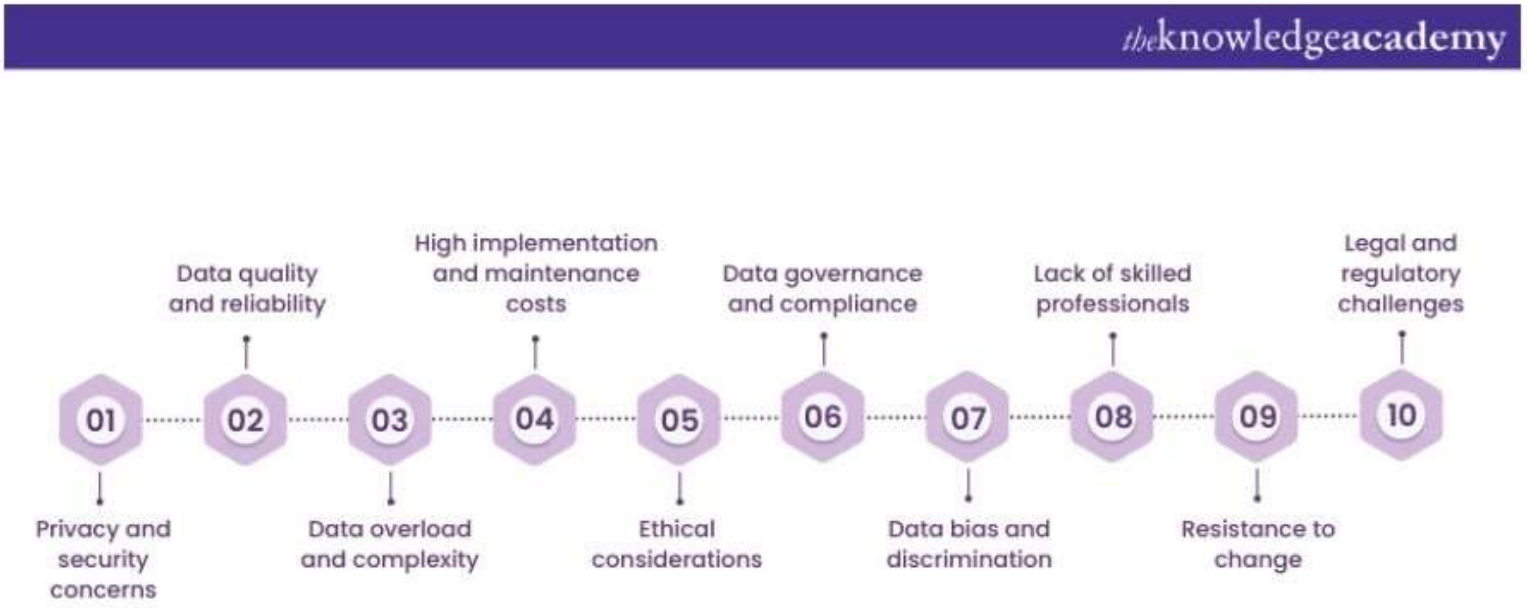
Advantages

- ✓ Enhanced decision-making
- ✓ Improved operational efficiency
- ✓ Personalisation and customer insights
- ✓ Cost savings
- ✓ Competitive advantage
- ✓ Innovation and product development

Disadvantages

Disadvantages of Big Data

Now, that we know its advantages, we will expand on the Disadvantages of Big Data.



Assessment 1

1. List out the advantages of big data Storage

a)_____

b)_____

c)_____

d)_____

2. Identify the disadvantages of big data Storage

a)_____

b)_____

c)_____

d)_____

REFERENCES

1. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann/Elsevier Publishers, 2013
3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.

THANK YOU