SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai
Accredited by NAAC-UGC with 'A++' Grade (Cycle III) &
Accredited by NBA (B.E - CSE, EEE, ECE, Mech & B.Tech.IT)
COIMBATORE-641 035, TAMIL NADU

DEPARTMENT OF COMPUTER APPLICATIONS

19CAE716 – DATA SCIENCE

UNIT – I: INTRODUCTION TO DATA SCIENCE

TOPIC: BIG DATA ECOSYSTEM AND DATA SCIENCE

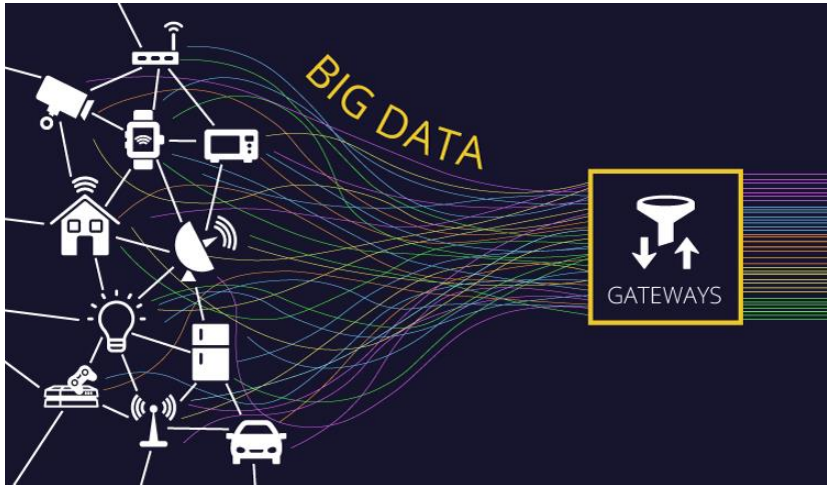
Introduction

In the digital age, the proliferation of data has reached unprecedented levels, giving rise to the need for innovative approaches to harness, analyze, and derive meaningful insights from vast datasets. This synergy between the expansive Big Data ecosystem and the analytical prowess of Data Science has become a cornerstone in the evolution of technology and business intelligence.

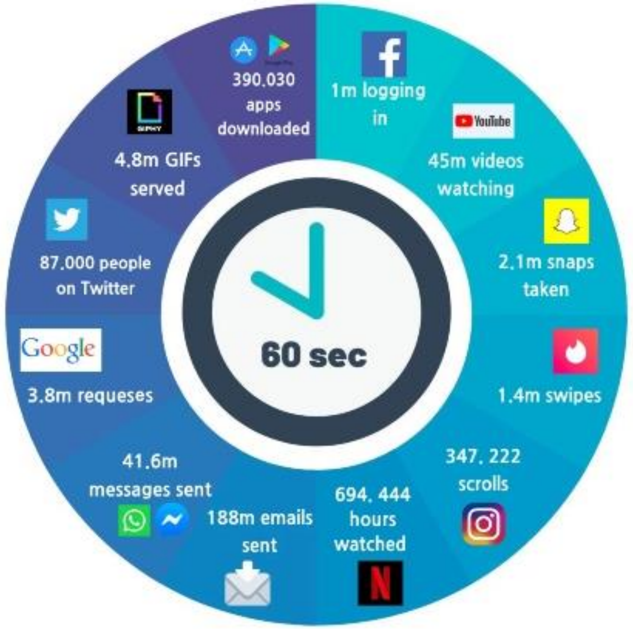
BIG DATA ECOSYSTEM

Data Generation & Ingestion

- ✓ The journey begins with the sheer volume of data generated daily.
- ✓ From the myriad of Internet of Things (IoT) devices to the constant flow of information on social media platforms, data is the lifeblood of the digital world.
- ✓ Technologies such as Apache Kafka and Flume play a pivotal role in ingesting and handling real-time data streams, ensuring that no data is lost in the deluge.

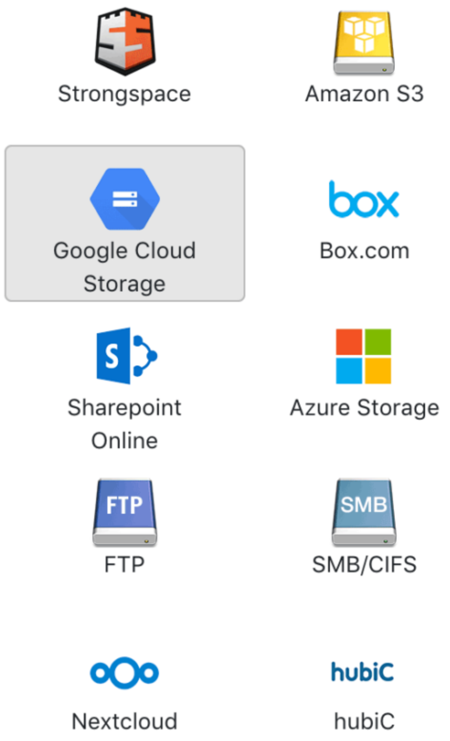


p
CSV
YOUR DATABASE
MySQL



Data Storage

- As data floods in, it needs a home.
- Enter the Hadoop Distributed File System (HDFS), a distributed file system designed to store massive amounts of data across clusters of computers.
- Cloud-based solutions like Amazon S3 and Google Cloud Storage have also become prominent players in this space, providing scalable and flexible storage options.



Data Processing

✓ Once stored, the data needs to be processed efficiently.

✓ MapReduce, with its parallel processing capabilities, and Apache Spark, a fast, in-memory data processing engine, are stalwarts in handling the complexities of large-scale data analytics.

✓ These technologies enable organizations to extract valuable insights from their data with unprecedented speed and accuracy.

Data Query and Analysis

- To interact with the stored data, tools like Hive and PrestoDB come into play.
- They provide a layer of abstraction, allowing users to query and analyze data using familiar SQL-like interfaces.
- This accessibility is crucial for data scientists and analysts to derive meaningful patterns and trends.

Data Visualization

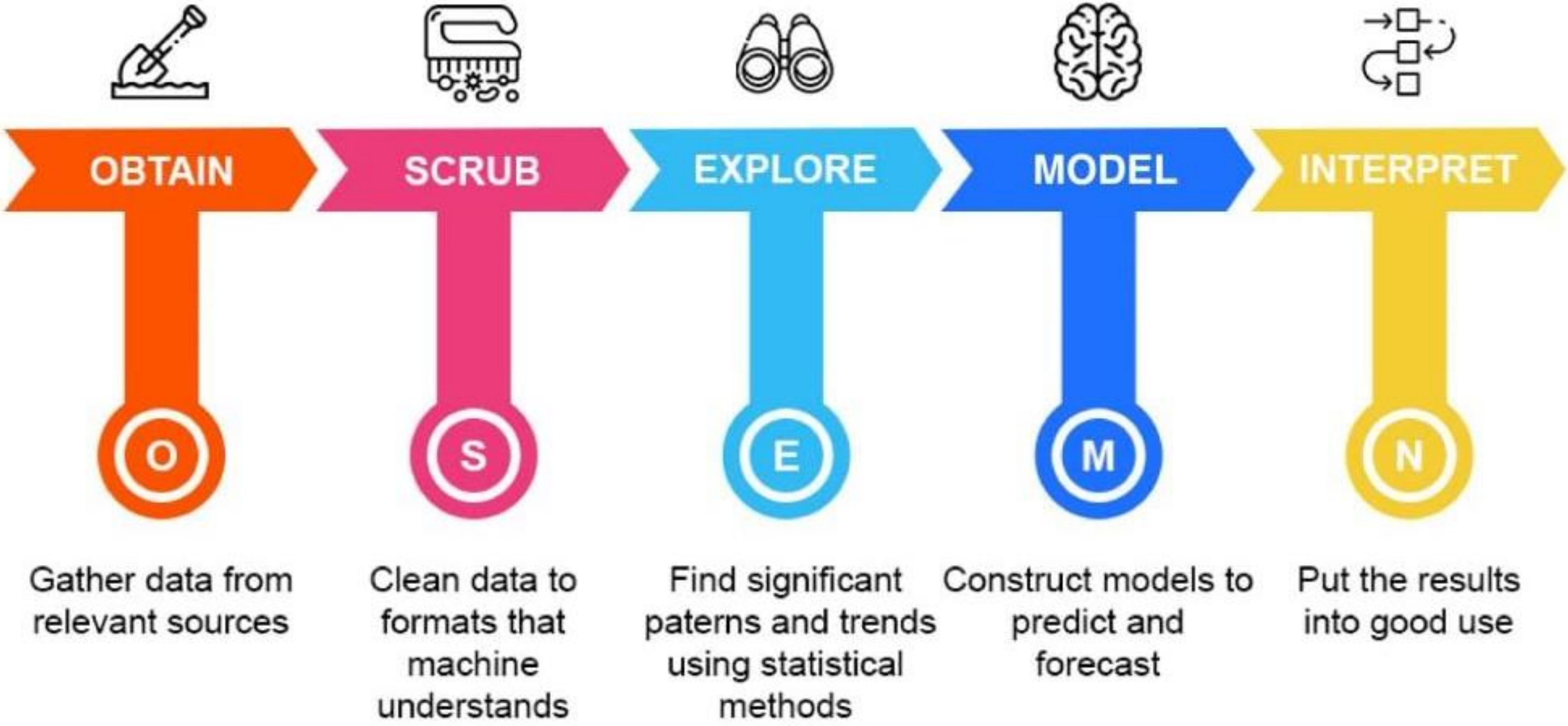
- ✓ The final step in the big data journey is making sense of the analyzed data.
- ✓ Visualization tools such as Tableau, Power BI, and D3.js empower users to create interactive and comprehensible visual representations of complex datasets.
- ✓ This visual storytelling is indispensable for conveying insights to stakeholders in a compelling manner.

+ableau

D3.js

Power BI

DATA SCIENCE PROCESS



Gather data from relevant sources

Clean data to formats that machine understands

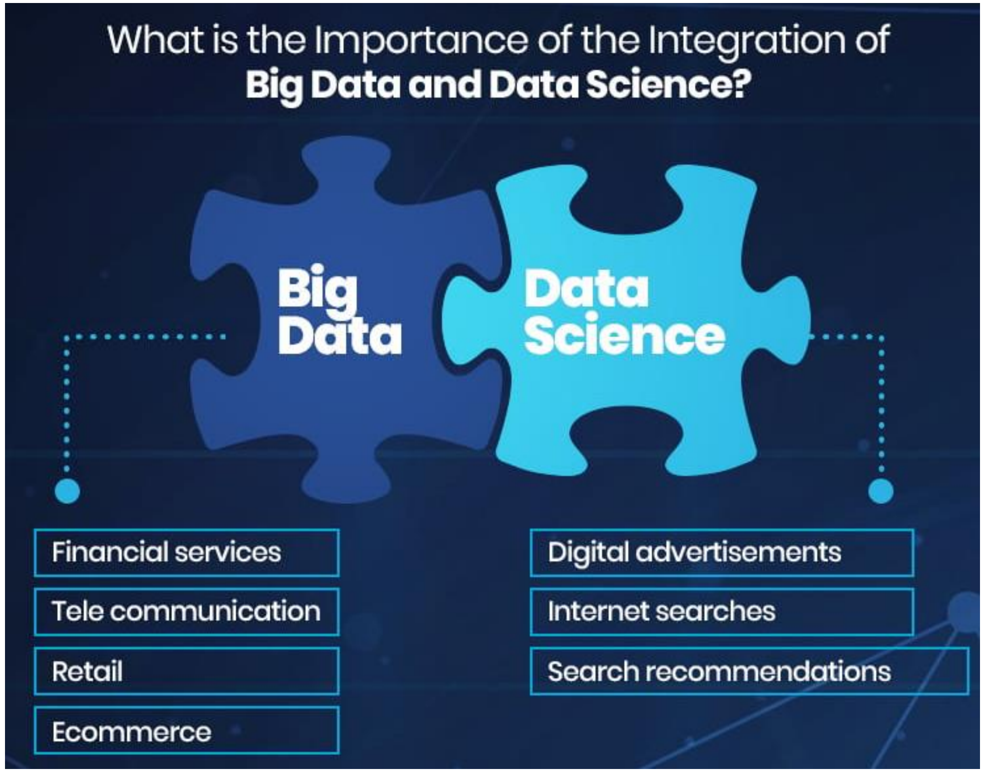
Find significant paterns and trends using statistical methods

Construct models to predict and forecast

Put the results into good use

Integration of DS and BD

- ✓ The integration of the Big Data ecosystem and Data Science represents a paradigm shift in how organizations approach data.
- ✓ It's a holistic approach that transforms data from being a challenge to an asset—a strategic resource that fuels innovation and competitiveness.
- ✓ As technology continues to evolve, this integration will play a pivotal role in shaping the future of data-driven decision-making and propelling organizations toward new frontiers of discovery and insight.


What is the Importance of the Integration of
Big Data and Data Science?
Big Data
Data Science
Financial services
Tele communication
Retail
Ecommerce
Digital advertisements
Internet searches
Search recommendations