DS598

ENGINEERING FOR BIG DATA WORKLOADS

-

BOSTON UNIVERSITY

FACULTY OF COMPUTING AND DATA SCIENCE

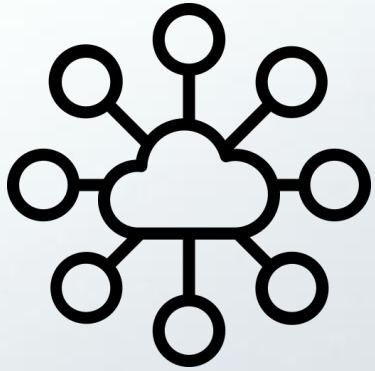
PROFESSOR CHRIS SEFERLIS, MBA

FALL 2024

LECTURE 2 – 09/10/2024

LECTURE 2 - BIG DATA FUNDAMENTALS

- Introduction to Big Data
- Characteristics of Big Data
- Big Data Lifecycle
- Big Data Ecosystem and Technologies
- Big Data Challenges & Opportunities




1
INTRODUCTION
TO BIG DATA

What is Big Data?

- Refers to datasets so large or complex that traditional data processing tools are inadequate to handle them.
- Big data is characterized by its ability to handle vast amounts of data generated from a variety of sources.

Importance of Big Data

- Big data is integral to modern analytics, AI, and real-time decision-making systems. It drives innovation, operational efficiencies, and competitive advantage in industries like finance, healthcare, e-commerce, and government.

2.2) Why Big Data Matters:

2.2) Why Big Data Matters

Global Data Growth

- The world is generating data at an unprecedented rate due to the rise of social media, IoT devices, and e-commerce platforms.

Characteristic	Amount per minute
USD traded in treasury bonds	398,000,000
Emails sent	241,000,000
WhatsApp messages sent	41,600,000
Global hours spent online	25,100,000
Searches on Google	6,300,000
Facebook posts liked	4,000,000
USD sent on Venmo	1,000,000
Reels sent via DM on Instagram	694,000
USD spent on Amazon	455,000
X (Twitter) posts sent	360,000
USD spent on DoorDash	122,000
Taylor Swift song streamed	69,400
Hours of content watched on Twitch	48,000
Chat GPT prompts sent	6,944
LinkedIn resumes submitted	6,060
Instagram Threads downloaded	3,720
Airbnb guest book stays	747
Data produced by the average person (MB)	102
Years of streaming content watched	43
DDOS attacks launched	30

https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

Economic Impact

- Organizations that can harness big data effectively gain strategic insights, identify trends, optimize operations, and innovate new products.

Real-World Examples

Social media analytics for sentiment analysis.

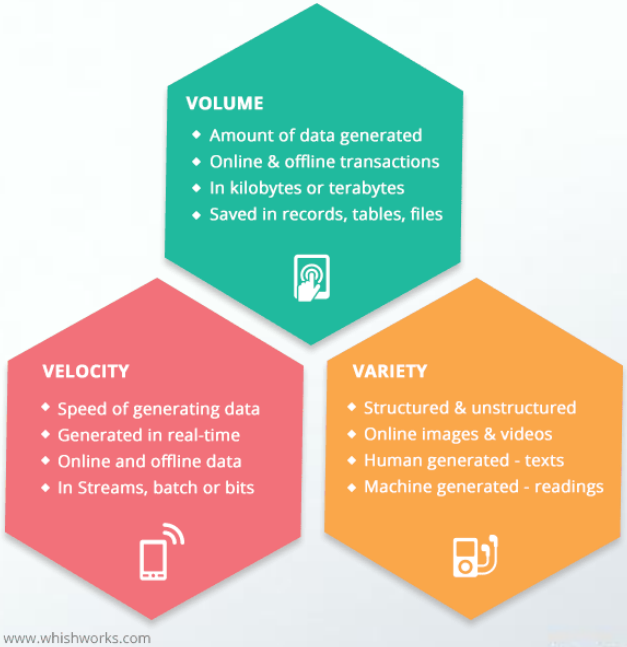
Predictive maintenance in manufacturing via IoT sensors.

Personalized recommendations in e-commerce.

2

CHARACTERISTICS OF BIG DATA

The 3Vs


THE 3Vs OF BIG DATA
VOLUME
Amount of data generated
Online & offline transactions
In kilobytes or terabytes
Saved in records, tables, files
VELOCITY
Speed of generating data
Generated in real-time
Online and offline data
In Streams, batch or bits
VARIETY
Structured & unstructured
Online images & videos
Human generated - texts
Machine generated - readings
www.whishworks.com

Volume

- **Definition:**
  - Refers to the massive amounts of data generated every second.
  - This could be structured (from relational databases), semi-structured (like JSON), or unstructured (videos, texts).
- **Example:**
  - Terabytes of data generated by social media platforms every day, data logs from autonomous vehicles, etc.

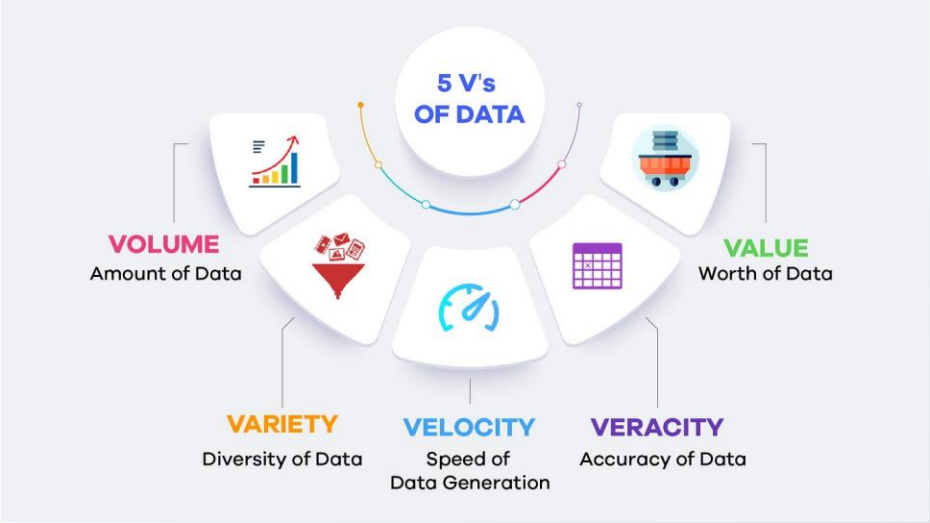
Velocity

- **Definition:**
  - Refers to the speed at which data is generated, processed, and analyzed.
  - Real-time data processing is becoming increasingly essential for applications such as fraud detection, algorithmic trading, and recommendation systems.
- **Examples:**
  - Real-time data streaming from IoT sensors or live transaction data in stock markets.

Variety

- **Definition:**
  - Describes the different types and formats of data being generated—structured, semi-structured, and unstructured.
  - Big data involves the integration and processing of these heterogeneous data types.
- **Examples:**
  - Combining relational database entries with multimedia files, log data, and social media feeds.

Beyond the 3Vs >> 5Vs



Veracity

- **Definition:**
  - Refers to the quality, accuracy, and trustworthiness of the data. Handling noisy, inconsistent, or incomplete data is one of the significant challenges in big data analytics.
- **Examples:**
  - Social media sentiment analysis must account for sarcasm, fake news, and misinformation.

Value

- **Definition:**
  - Refers to the actionable insights and economic value derived from big data.
  - Not all big data is useful; the challenge lies in identifying the valuable data and using it effectively.
- **Examples:**
  - E-commerce platforms derive value by predicting customer preferences and personalizing the shopping experience, based on user behavior data.

Group Digression

20 Minutes

Read and prepare to discuss

- Big data in healthcare: Prospects, challenges and resolutions | IEEE Conference Publication | IEEE Xplore

Review questions

- How has the integration of ICT and big data transformed healthcare in terms of patient care and cost reduction?
- What are the main challenges of implementing big data in healthcare, particularly regarding security and privacy?
- In what ways do the '5 Vs' of big data (volume, velocity, variety, variability, veracity) present specific challenges to healthcare providers?
- How can healthcare institutions overcome resistance to change when adopting big data solutions?
- What are the ethical considerations surrounding big data in healthcare, especially regarding patient privacy and data ownership?

10 MINUTE BREAK


3

BIG DATA LIFECYCLE

3.1) Data Generation

Traditional Data Generation

- Data is generated from standard sources like transactional databases, small-scale user interactions, or file systems.
- Typically, structured data, generated at predictable rates.

Big Data Generation

- **Difference**: The volume and velocity are significantly higher.
- **Big Data Sources**: IoT devices, social media, sensors, web logs, satellite imagery, and real-time interactions.
- **Challenge**: Managing the continuous flow of massive data streams from diverse sources (unstructured, semi-structured, structured). High-frequency data needs to be captured in real-time.

3.2) Data Ingestion

Traditional Data Ingestion

- Data is collected periodically (batch ingestion) and transferred into databases or warehouses.
- Manual or semi-automated ingestion is typical.



Big Data Ingestion

- **Difference**: Ingestion must support real-time data streaming, as well as batch ingestion for large datasets.
- **Big Data Ingestion Tools**: Apache Kafka, Apache Flume, Amazon Kinesis, and Apache Spark for streaming.
- **Challenge**: Handling both real-time streaming and large-scale batch processes efficiently. Ensuring the system scales to handle peaks in data volume without losing data.

3.3) Data Storage and Management

Traditional Data Management

- Stored in structured formats, often in relational databases (SQL), with clear data models.
- Data stored on-premises or in small-scale cloud environments.



Big Data Storage & Management

- **Difference**: Distributed storage is a must due to the sheer volume of data.
- **Big Data Storage Technologies**: Hadoop Distributed File System (HDFS), cloud storage (AWS S3, Google Cloud Storage), NoSQL databases.

Big Data Storage & Management

- **Challenge**: Ensuring scalability, fault tolerance, and high availability. Data is often stored across multiple locations, which introduces complexity in terms of synchronization and consistency.
- **Data Governance**: Ensuring data security, quality, and compliance becomes more difficult as the data grows in volume and variety.

3.4) Data Processing and Analysis

Traditional Processing & Analysis

- Processed using traditional batch methods, often with SQL queries, spreadsheets, or data warehouse reporting tools.
- Data analysis tools typically focus on descriptive analytics.

Big Data Processing & Analysis

- **Difference**: Requires both real-time and batch processing capabilities.
- **Big Data Processing Frameworks**: Apache Hadoop with MapReduce (for batch processing), Apache Spark (for in-memory and real-time processing),.

Big Data Processing & Analysis

- **Challenge**: Processing petabytes of data in a timely manner. Real-time analytics require in-memory processing, while batch analytics handle large historical datasets.
- **Advanced Analytics**: Big data allows for predictive analytics, machine learning, and deep learning at scale, going beyond traditional descriptive analytics.

3.5) Data Visualization

Traditional Data Visualization

- Data is visualized through basic reporting tools (e.g., Reporting Services, Excel, Tableau, Cognos) using pre-aggregated or small datasets.
- Relatively simple dashboards and charts are enough for decision-making.



Big Data Visualization

- **Difference**: Visualization must handle complex, massive datasets in real-time, requiring interactive dashboards.
- **Big Data Visualization Tools**: Power BI, Tableau (for big data), custom-built visualization layers on top of big data platforms like Apache Zeppelin or Kibana for real-time streaming data.


Big Data Visualization

- **Challenge**: Presenting insights from high-velocity, high-volume data streams while ensuring scalability in the visualization tools themselves. Handling unstructured data visualizations (e.g., images, videos, logs) becomes more complex.

3.6) Data Archival and Disposal

Traditional Archival & Disposal

- Data is archived according to basic retention policies, often stored on-premises, off-site storage, or in cloud backups for regulatory compliance.
- Disposal is straightforward, with scheduled deletion or destruction after certain periods.

Big Data Archival & Disposal

- **Difference**: Requires long-term storage solutions capable of scaling with continuous data generation.
- **Big Data Archival**: Cold storage solutions (e.g., Amazon Glacier, Azure DataBox) for large, infrequently accessed datasets.



Big Data Archival & Disposal

- **Challenge**: Balancing cost efficiency and accessibility for archived data. Implementing smart archival solutions that minimize storage costs while allowing for retrieval when necessary. Ensuring proper disposal without violating data privacy laws (like GDPR, HIPAA) in large-scale environments.

4

Big Data Ecosystem & Technologies

4.1) Distributed Storage Systems

Hadoop Distributed File System (HDFS)

- **Definition:** A distributed storage system designed for storing vast amounts of data across multiple servers. It enables fault tolerance by replicating data across nodes.
- **Importance**: A backbone for large-scale data storage, ensuring scalability and reliability.
- We will get into detail on Hadoop later in the semester

Cloud Storage

- **Definition:** Cloud platforms (AWS S3, Google Cloud Storage, Azure Blob) provide scalable and flexible storage solutions, enabling organizations to store and access data from anywhere.
- **Importance**: Allows on-demand storage and eliminates the need for costly on-premises infrastructure.

4.2) Data Processing Frameworks

MapReduce

- **Definition:** A programming model for processing large datasets by breaking down tasks into smaller, parallel jobs.
- **Importance**: Ideal for batch processing of big data, simplifying large-scale computations.

Apache Spark

- **Definition:** An open-source data processing engine that supports both batch and real-time data analytics, known for its in-memory processing.
- **Importance**: Faster than MapReduce due to its ability to process data in-memory, and supports complex tasks like machine learning and graph processing.

4.3) Real-Time Streaming Technologies

Kafka/Storm/Flink

- **Definition:** Streaming platforms that enable real-time data ingestion and processing, allowing organizations to act on data as it is generated.
- **Use Cases**: Fraud detection, real-time personalization, predictive maintenance.

4.4) Data Integration and Management Tools

ETL Pipelines

- **Definition:** Extract, Transform, Load (ETL) processes used to integrate data from different sources into a centralized data repository.
- **Tools**: Informatica, Talend, Apache NiFi.

Data Governance

- **Definition:** Ensures that data is managed securely, with compliance to privacy and security standards.
- **Example**: Ensuring GDPR compliance in customer data processing.

5

CHALLENGES & OPPORTUNITIES

5.1) Challenges

Scalability

- **Challenge**: Handling petabytes or exabytes of data requires robust and scalable infrastructure that can grow with the data.
- **Solutions**: Distributed computing, cloud infrastructure, and scalable storage solutions help manage growing data volumes.

Data Integration

- **Challenge**: Integrating data from multiple sources, in different formats, and of varying quality.
- **Solutions**: Use of ETL processes, data lakes, and integration platforms like Talend or Apache Nifi.



Data Quality and Consistency

- **Challenge**: With the variety of data sources, ensuring high-quality, consistent, and reliable data becomes a significant challenge.
- **Solutions**: Data cleaning, validation, and governance practices are critical to ensuring that data is useful and trustworthy.



Security and Privacy Concerns

- **Challenge**: With growing amounts of data, especially personal and sensitive information, security breaches and data privacy violations are major risks.
- **Solutions**: Encryption, anonymization, and compliance with regulations (e.g., GDPR, HIPAA) are necessary to secure big data environments.



Real-Time Processing

- **Challenge**: Real-time data streams require low-latency processing, which can be difficult to achieve with traditional systems.
- **Solutions**: Implementing in-memory processing (e.g., Apache Spark) and stream processing platforms (e.g., Apache Kafka).

5.2) Opportunities

Business Innovation and Optimization

- Big data allows businesses to innovate by deriving insights from large datasets that were previously inaccessible or too costly to analyze.

- **Customer Personalization**: Using big data to tailor product recommendations, improve customer experience, and optimize marketing strategies.
- **Operational Efficiency**: Data-driven decisions allow businesses to optimize supply chains, reduce waste, and improve processes.

Predictive Analytics and AI

- The ability to predict future trends based on historical data is a significant advantage of big data analytics.
- Machine learning models thrive on large datasets, enabling businesses to predict customer behavior, product demand, and more.
- **Predictive Maintenance**: Using sensor data to predict equipment failures and schedule maintenance before breakdowns occur.
- **Financial Forecasting**: Using past financial performance to model future outcomes and trends in markets.

Healthcare and Life Sciences

- Big data has transformative potential in healthcare by providing personalized medicine, improving diagnostic accuracy, and advancing drug discovery.
- **Precision Medicine**: Using genetic, environmental, and lifestyle data to provide more accurate diagnoses and personalized treatment plans.
- **Epidemic Prediction**: Analyzing public health data to predict outbreaks and respond to health crises in real-time.

Smart Cities and IoT

- Big data enables cities to improve urban planning, reduce traffic congestion, and manage energy more efficiently.
- **Traffic Management**: Analyzing real-time traffic data to optimize traffic light timing, reduce congestion, and improve public transport systems.
- **Energy Efficiency**: Monitoring energy consumption in real-time to optimize usage and reduce waste.

5.3) Career Opportunities

Emerging Roles

- **Data Engineer**: Focuses on building and maintaining big data pipelines.
- **Data Scientist**: Extracts insights from big data using statistical and machine learning techniques.
- **Big Data Architect**: Designs the architecture for big data solutions, ensuring scalability and performance.

Skills Needed

- **Technical Skills**: Proficiency in programming (Python, Java), knowledge of big data tools (Hadoop, Spark), data modeling, and SQL.
- **Analytical Skills**: Ability to interpret data, understand trends, and make data-driven decisions.
- **Soft Skills**: Communication, teamwork, problem-solving, and a continuous learning mindset.

8

Big Data Chat

Article 1

Today's Article – 09.10.24

- **Topic: Big data in healthcare: Prospects, challenges and resolutions**
  - Iroju Olaronke; Ojerinde Oluwaseun
  - https://ieeexplore.ieee.org/abstract/document/7821747

- Presenters:
  - **https://pickerwheel.com/rng?id=LP6h7**

Big Data Chat Article 2 – 09.17.24

- **Big Data Architectures : A detailed and application oriented review**
  - Godson Koffi Kalipe, Rajat Kumar Behera
  - https://www.researchgate.net/publication/336915402_Big_Data_Architectures_A_detailed_and_application_oriented_review
- Please make sure you read the article and come prepared for a discussion in next week's lecture!

LET'S GO
DATA WARRIORS!

Thanks!

Today’s Reminder >>

- Homework 1 - Due
- Homework 2 – Discussion
- Check Project Team - Discussion
- Project Challenge 0 - Discussion