7BDIN006W
Big Data Theory and Practice

Lecture 1

Module Introduction. Big Data Definition. Features of Big Data.
Big Data Analytics. Ethics and Risks.

UNIVERSITY OF
WESTMINSTER

Welcome
to the BDT&P Module!

Module Leader

Dr Natalia Yerashenia
Lecturer in Data Science & Analytics

n.yerashenia3@westminster.ac.uk

Office 7.115, Cavendish Campus

Office Hours:
Tue 17:00–18:00; Wed 16:30–18:00

UoW Profile Link

Module Organisation

- Module Duration: One Semester, 12 Weeks
- Weekly Lectures (1.5hrs) & Tutorials (1.5hrs)
- Attendance is essential
- Lectures will be recorded. No tutorials recordings
- Practical Tasks: Oracle APEX, Neo4j, and more

Oracle APEX

!MODULE ASSESSMENT!

- 3 Assessments: ICT, Group Coursework, Presentation
- **ICT** – Week 8. In-Class Test. Duration: 60 minutes. Multiple choice questions.
- **Group CW** – Technical Report. Deadline: January 8, 2024
- **Presentation** – 5 minutes. Deadline: December 8, 2023
- Overall Module Mark = 30%ICT1+60%CW+10%P

Module Syllabus

- Foundations of Data Systems
- SQL and RDBMS
- NoSQL Databases
- Web and Data
- Big Data Tools and Frameworks
- Advanced Data Processing
- Data Quality and Governance
- Risk and Ethics in Data Management

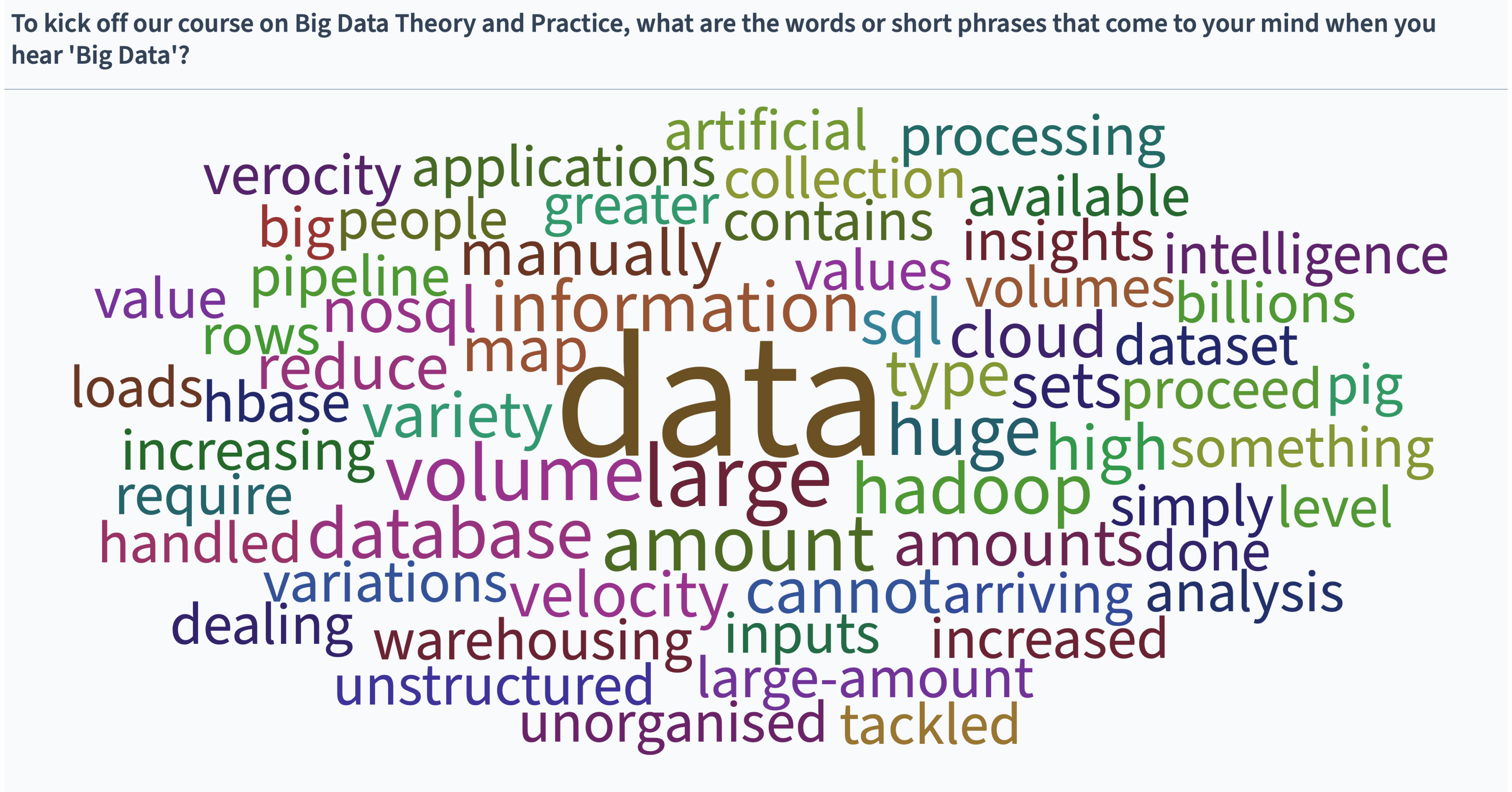
ACTIVITY: Word Cloud

Question: what are the words or short phrases come to your mind when you hear ‘Big Data” term?

NB. Think about its applications, challenges, tools, outcomes, or anything else you associate with

PollEv.com/nataliayerashenia404

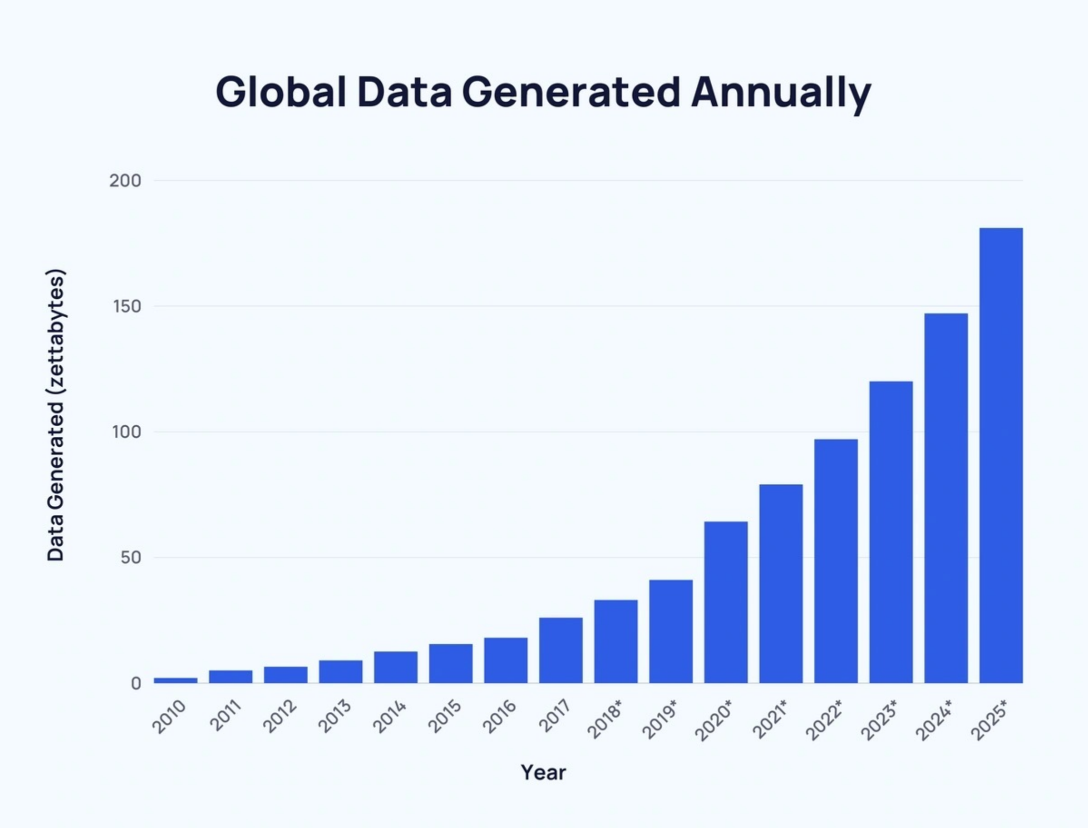
Poll Results



To kick off our course on Big Data Theory and Practice, what are the words or short phrases that come to your mind when you hear 'Big Data'?

Sep 2023

How much Data is generated every day?



According to the latest estimates, **328.77 million** terabytes of data are created each day.

90% of the world's data was generated in the last two years alone

The **120 zettabytes** generated in 2023 are expected to increase by **over 150%** in 2025, hitting **181 zettabytes**.

Big Data Definition

Big Data refers to extremely large datasets that are *difficult to manage and process* using traditional data management tools due to their volume, velocity, variety, and sometimes, veracity.

Big Data vs Traditional Data

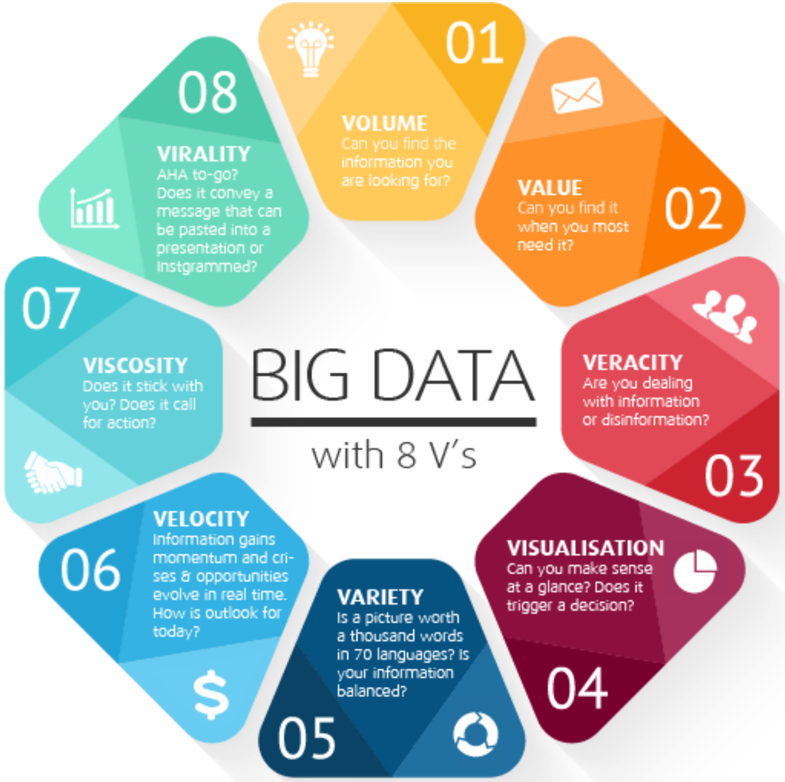
Aspect	Traditional Data	Big Data
Scale	Smaller in scale	Vast volumes requiring distributed storage solutions
	Managed with conventional DBMS	Necessitates specialized handling
Processing	Processed using standard DBMS tools	Utilizes specialized tools (e.g., Hadoop, Spark)
	Standard processing methods	Requires advanced, parallel processing for real-time analytics
Structure	Primarily structured (rows & columns)	Comprises structured, semi-structured, and unstructured data
	Easily organized & managed	Requires versatile storage & handling solutions
Insights	Yields descriptive & diagnostic insights	Offers predictive & prescriptive insights based on complex data patterns
	Focuses on “What happened?” and “Why?”	Provides foresight and actionable recommendations

A Brief History of Big Data

- **Before 2000s:** Development of relational databases; the advent of the internet.
- **2001:** Doug Laney defines Big Data as Volume, Velocity, and Variety.
- **2004-2006:** Google introduced MapReduce; Apache Hadoop is created.
- **2010s:** Expansion of Big Data technologies; the rise of Machine Learning and AI; increased focus on data privacy (e.g., GDPR).
- **2020s to Present:** Evolution and integration with cloud and edge computing; growing emphasis on real-time processing and ethical use of data.

Big Data has revolutionised the way we collect, store, analyse, and leverage data across various fields, driving advancements and raising ethical considerations.

Features of Big data



BIG DATA

with 8 V's

01 **VOLUME**
Can you find the information you are looking for?

02 **VALUE**
Can you find it when you most need it?

03 **VERACITY**
Are you dealing with information or disinformation?

04 **VISUALISATION**
Can you make sense at a glance? Does it trigger a decision?

05 **VARIETY**
Is a picture worth a thousand words in 70 languages? Is your information balanced?

06 **VELOCITY**
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?

07 **VISCOSITY**
Does it stick with you? Does it call for action?

08 **VIRALITY**
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

1.VOLUME

Vast amounts of data are generated every second from various sources like social media, transactions, sensors, etc.

2.VELOCITY

The *speed* at which new data is generated and collected.

3.VARIETY

Different *types of data*, including structured, semi-structured, and unstructured.

4.VERACITY

The quality and *reliability of the data*, addressing uncertainties due to inconsistency and availability.

5.VALUE

The *usefulness* of the data in creating economic value and enabling insights.

Big Data Technologies

Hadoop:

Open-source framework for distributed storage and processing of large datasets. It divides large files into smaller blocks (typically 128 MB or 256 MB) and stores multiple copies of these blocks across the cluster nodes to ensure fault tolerance.

MapReduce:

A programming model and processing engine for *parallel computat*ion of large datasets. It works in two phases: the Map phase (which filters and sorts data) and the Reduce phase (which performs summary operations).

Apache Spark:

An open-source, distributed computing system that provides a *fast and general-purpose cluster-computing framework* for big data processing. It was developed in response to limitations in the Hadoop MapReduce computing model, aiming to offer faster computation and more flexibility for various data processing tasks.

NoSQL Databases:

Databases are designed for storage and retrieval of *unstructured data*. Examples include MongoDB, Cassandra, Neo4j, and Couchbase. Neo4j is especially notable for its graph database, which is optimal for storing and querying highly interconnected data.

Cloud Computing:

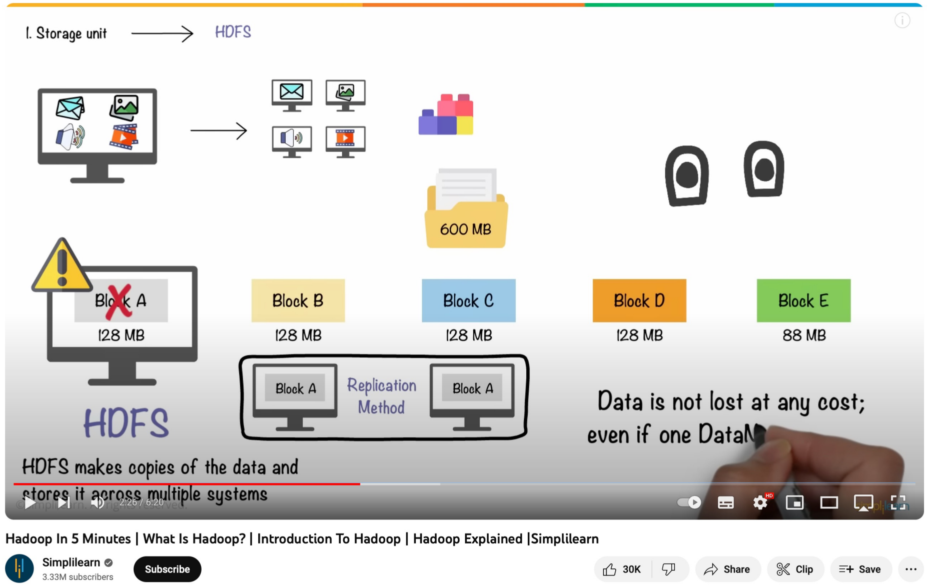
Provides on-demand computing resources over the internet. Enables scalable and *cost-effective Big Data storage* and processing. Examples include AWS, Google Cloud, and Azure.

Data Warehousing:

A *centralised repository* for storing large volumes of data from multiple sources. Examples include Amazon Redshift, Snowflake, and Google BigQuery.

A bit more details about Hadoop…

Hadoop In 5 Minutes*



HDFS makes copies of the data and stores it across multiple systems

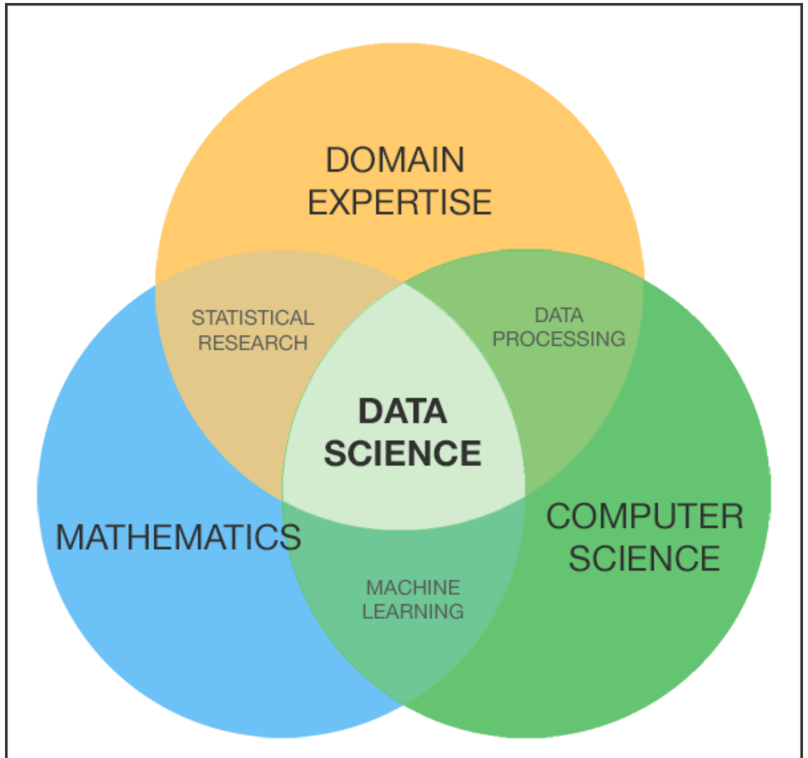
Hadoop In 5 Minutes | What Is Hadoop? | Introduction To Hadoop | Hadoop Explained |Simplilearn

Simplilearn
3.33M subscribers

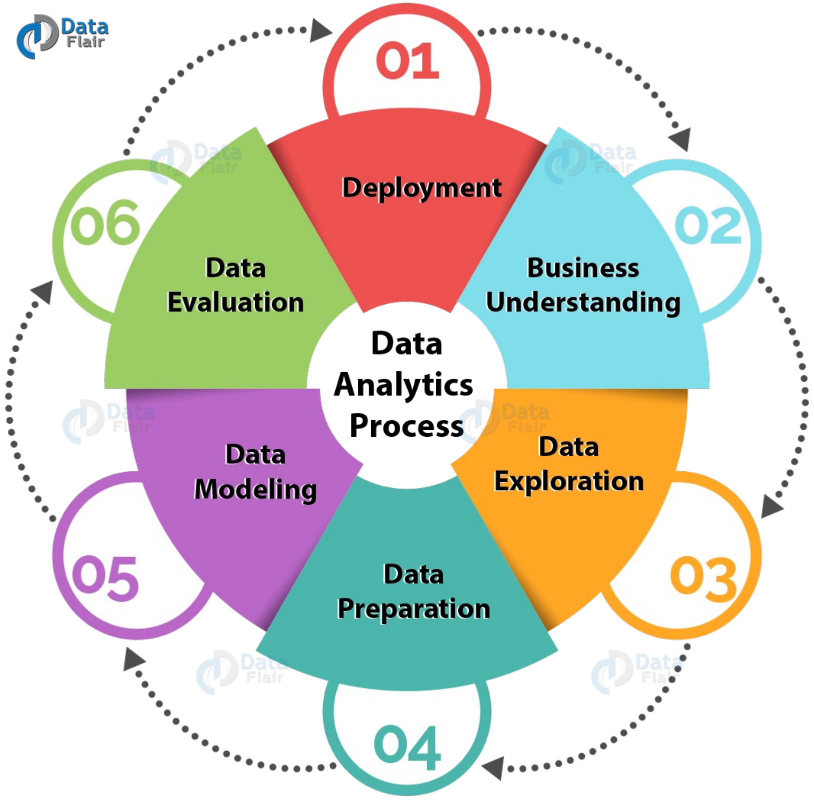
**Follow the link to watch a YouTube Video*

Big Data Analytics

Big Data Analytics involves examining, cleaning, transforming, and modelling large datasets to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information.



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.



Big Data Analytics Use Cases

- **Customer Behaviour Analysis:**
  Understanding customer preferences and improving experiences.
- **Fraud Detection:**
  Identifying unusual patterns and preventing fraudulent activities.
- **Supply Chain Optimization:**
  Enhancing efficiency and reducing costs in supply chains.
- **Healthcare Analytics:**
  Predicting disease outbreaks and improving patient care.
- **Sentiment Analysis:**
  Analysing public opinion and social media trends.

Ethics & Risks in Big Data

1. **Privacy:**
   Safeguarding individuals' data and respecting their privacy preferences.
2. **Consent:**
   Ensuring informed consent for data collection and use.
3. **Transparency:**
   Clear communication about how data is collected, processed, and used.
4. **Bias & Fairness:**
   Addressing biases in data and algorithms to prevent discrimination.
5. **Data Ownership:**
   Establishing and respecting ownership and rights to data.

1. **Security Breaches:**
   Protecting data from unauthorised access and cyber-attacks.
2. **Data Misuse:**
   Preventing the inappropriate use of data for malicious purposes.
3. **Quality & Accuracy:**
   Ensuring the reliability and accuracy of data and insights derived.
4. **Regulatory Compliance:**
   Adhering to local, national, and international data protection laws.
5. **Reputation Damage:**
   Managing the risk of harm to organisational reputation due to unethical practices.