University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Big Data Tools and Techniques

Week 1

Big Data

2025

Expectations

1. Choose a quiet place to attend the class and please concentrate during the lecture.
2. Put your questions in Padlet and I will review them in the due time (Padlet link is in BB, week 1, Lecture folder for Q&A week1).
3. You can find a handout on BB.
4. We will have 5 mins break after the first hour of the lecture (please remind me).
5. Jisc code will be shared during the break time.

Learning Outcomes

1. To explain the significance and applications of various analytical goals and types.
2. To comprehend the characteristics and challenges associated with big data.
3. To describe how the properties of big data contribute to its overall nature.
4. To understand the big data analytics lifecycle for real world projects.
5. To define cloud computing.

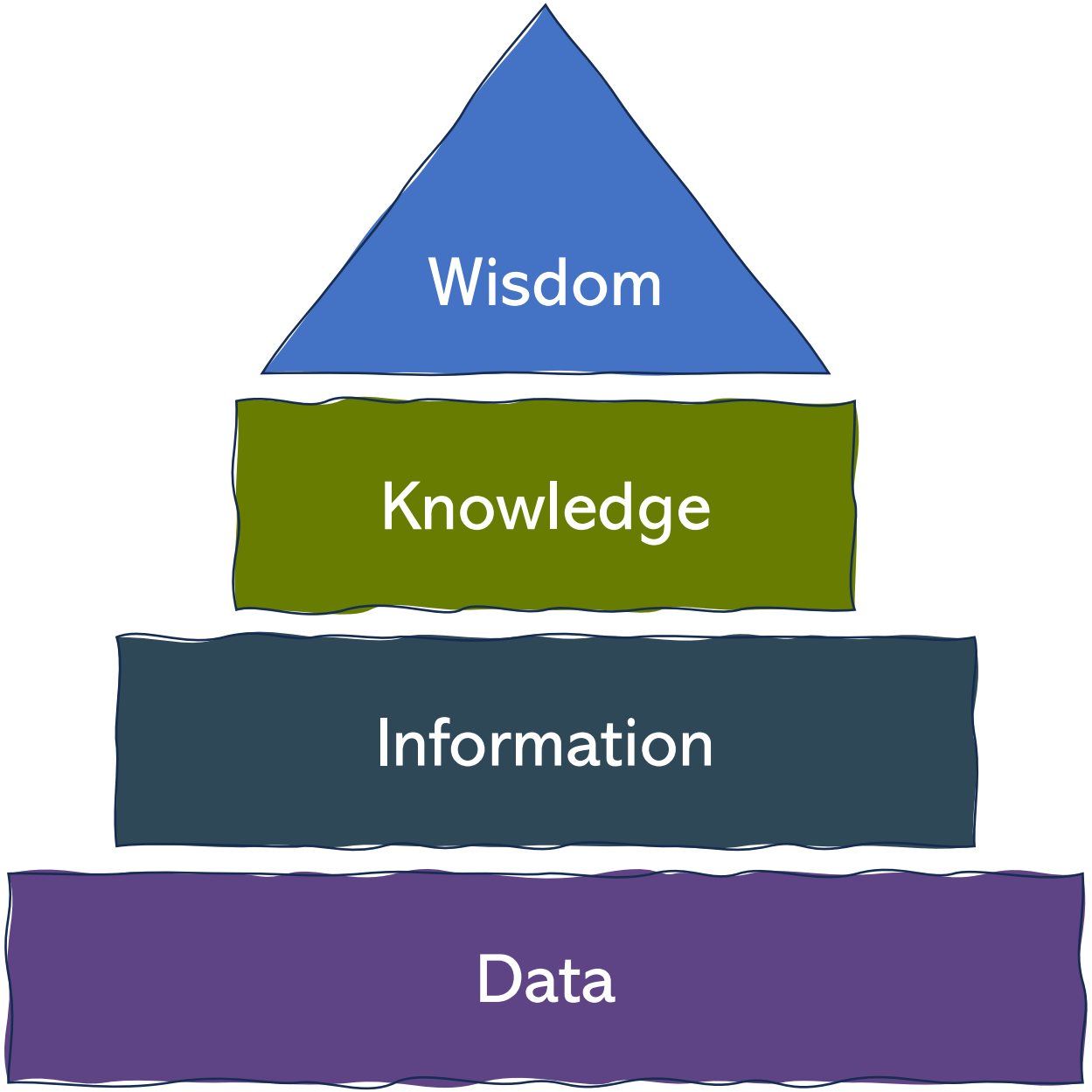
BDTT

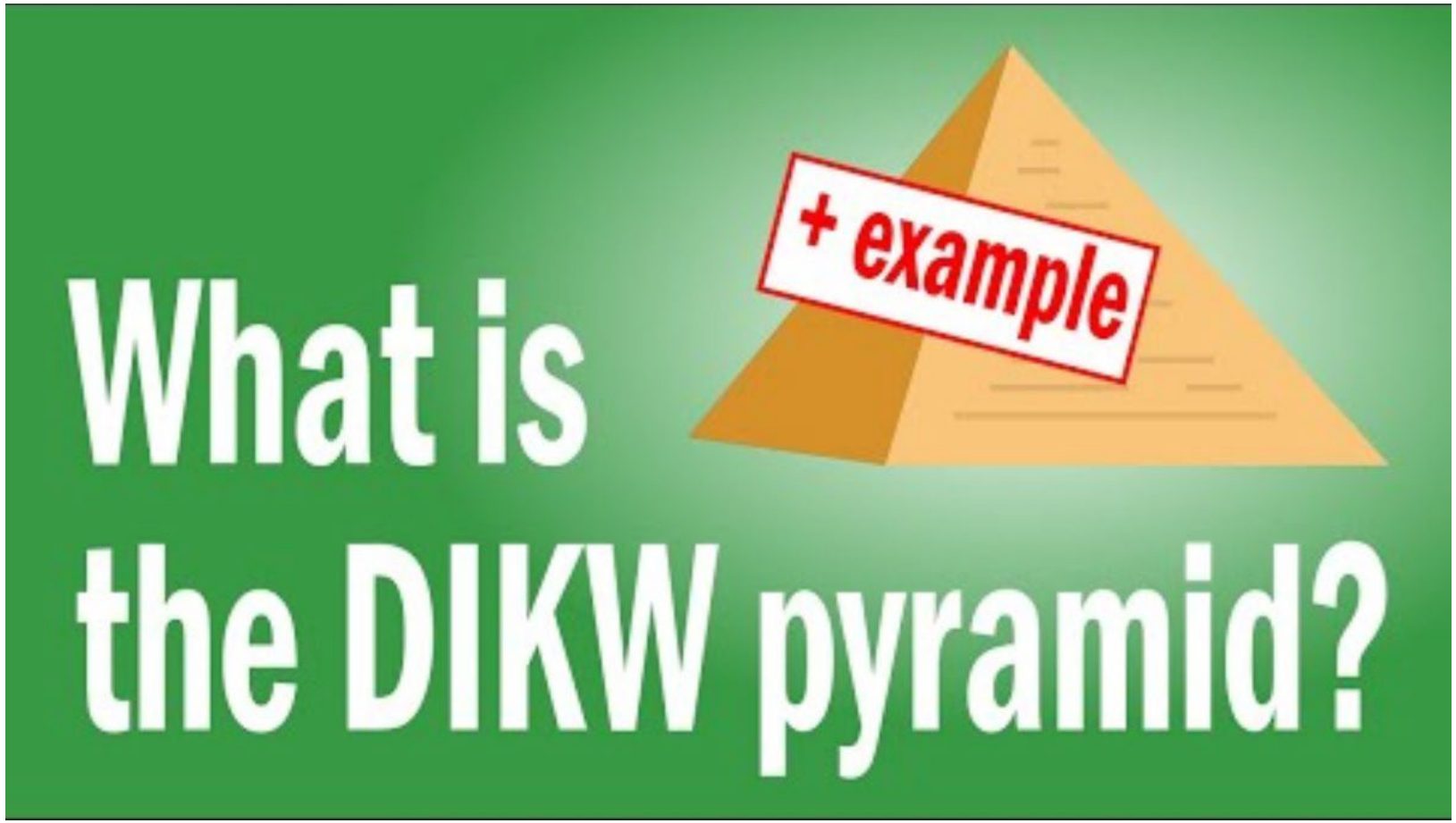
BDTT is very hands on.

- **Lectures**: about 2hrs lectures.
- **Workshops** and **Tutorials**: 3hrs you work through the labs.
- **Drop-in** sessions: weekly 1hr starting from week 3 (attendance is optional).

Assignment, given out in week 4.

Introduction


Wisdom
Knowledge
Information
Data

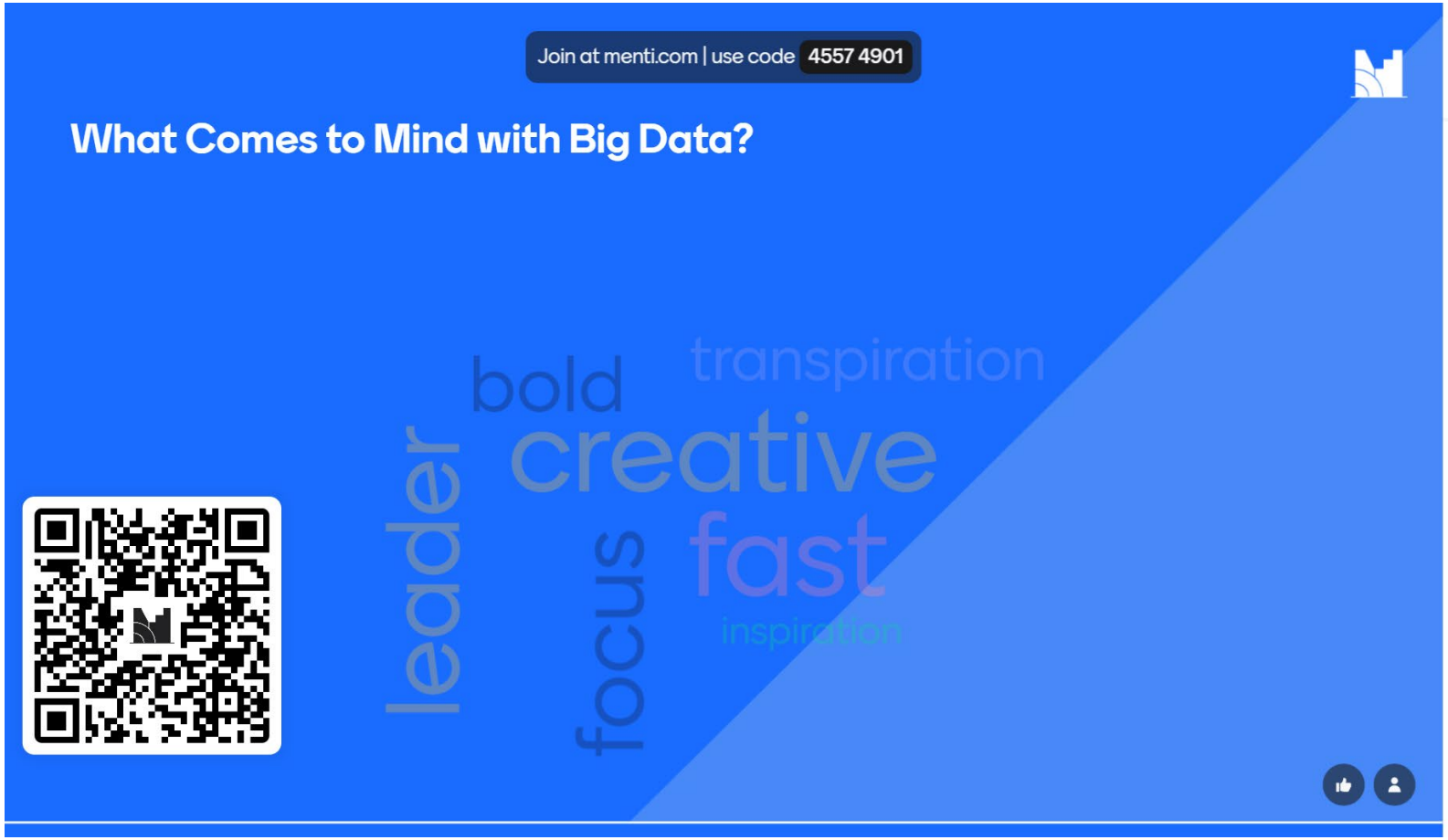

What is
the DIKW pyramid?
+ example

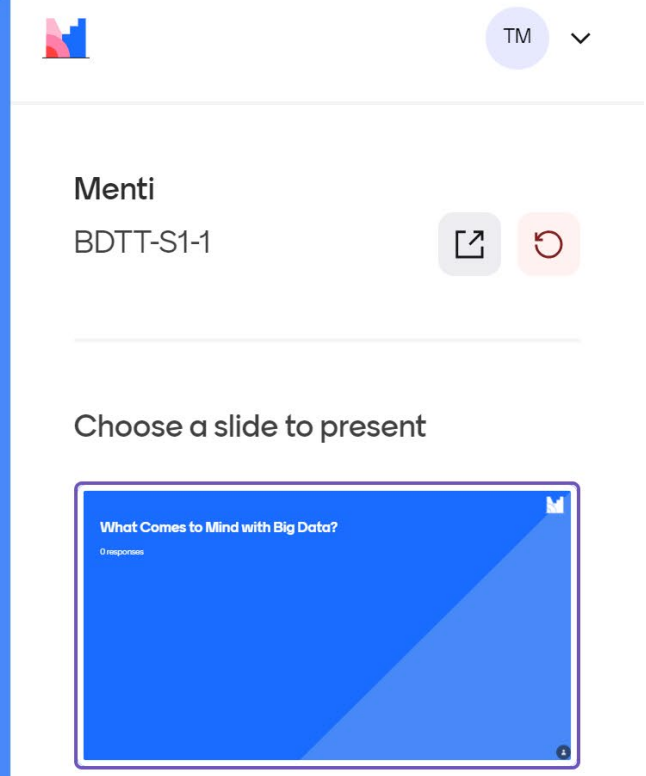
Different Types of Analysis

Types of Analysis	Questions	Value	Complexity	Example
Descriptive	What happened?	Summarizes past performance.	Low	Monthly sales report.
Diagnostic	Why did it happen?	Identifies causes of past events.	Moderate	Analysing drop in sales for a region.
Predictive	What will happen?	Forecasts future outcomes.	High	Predicting product demand for the next quarter.
Prescriptive	What should we do about it?	Recommends actions to achieve goals.	Very high	Optimizing inventory levels for the predicted demand to avoid stockouts.

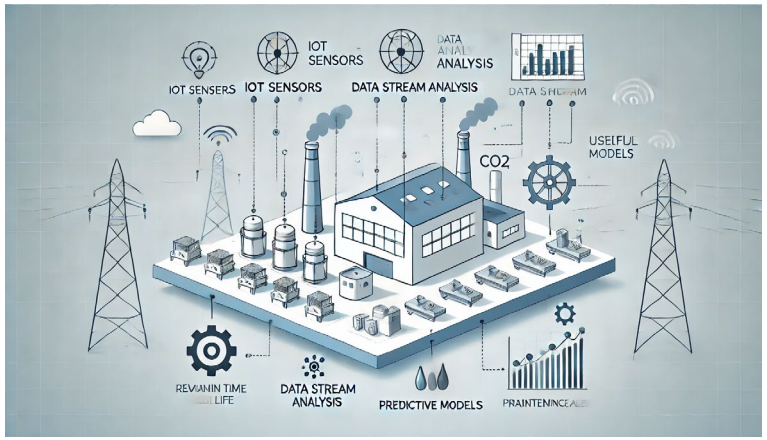
What is Big Data?

Activity 1

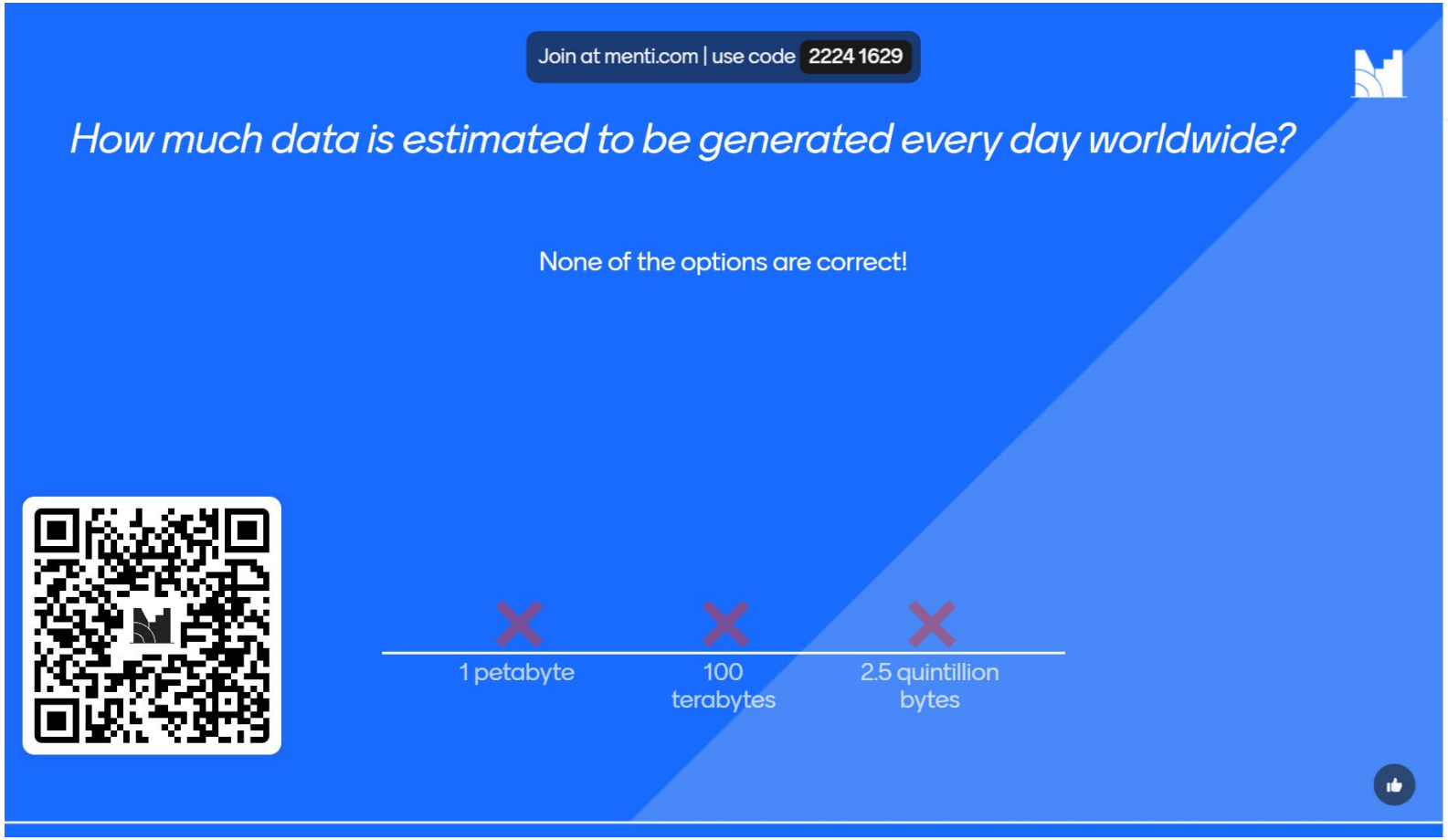

Join at menti.com | use code 4557 4901
What Comes to Mind with Big Data?
bold
transpiration
leader
creative
focus
fast
inspiration

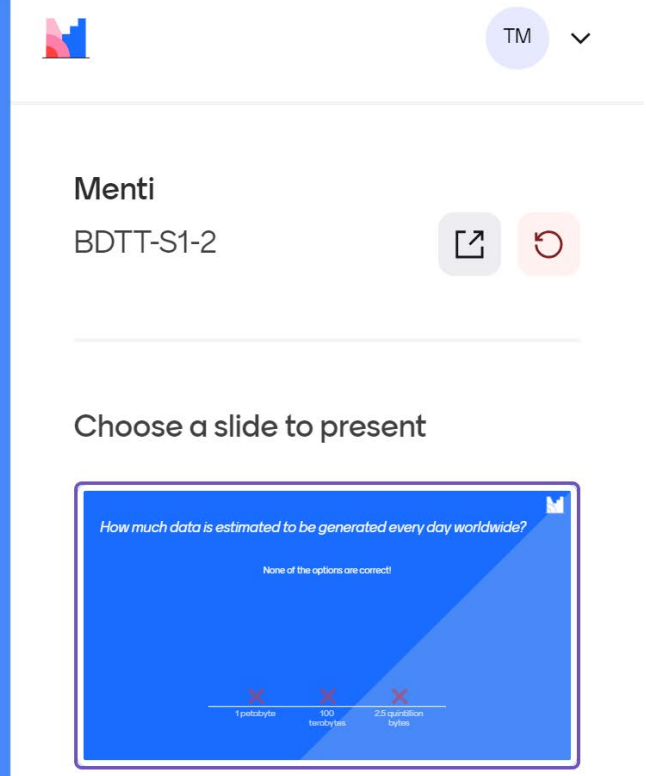

TM
Menti
BDTT-S1-1
Choose a slide to present
What Comes to Mind with Big Data?

Big Data: A few examples



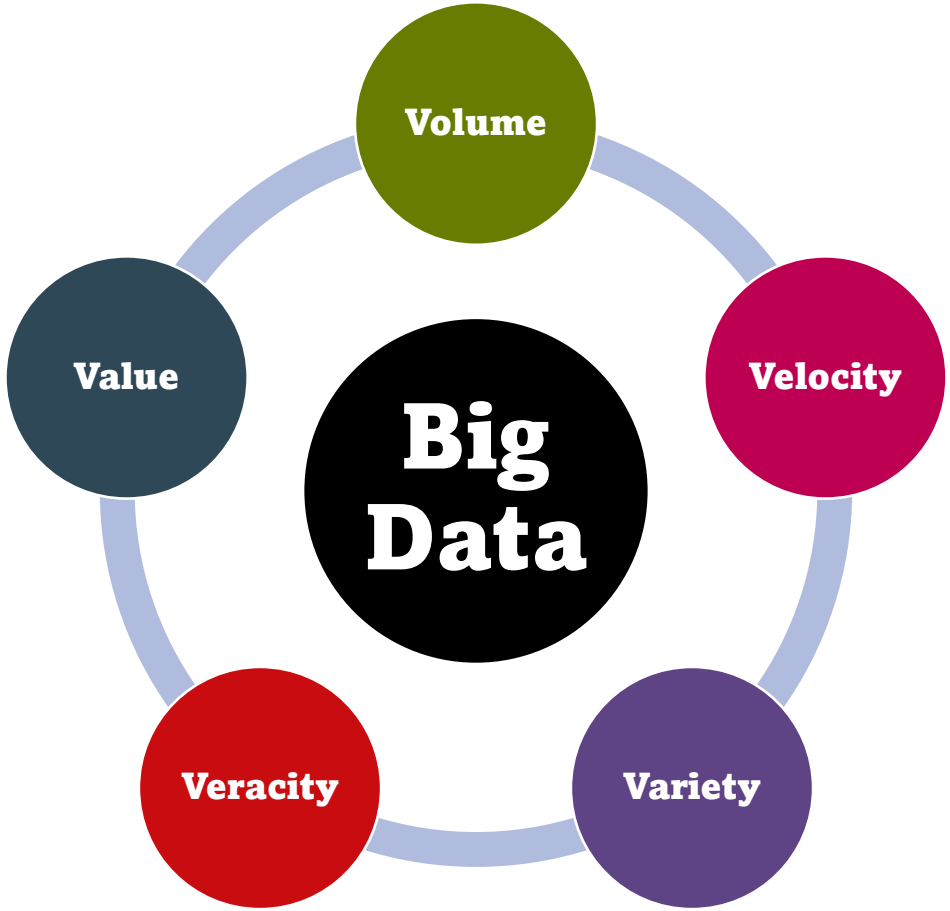
Activity 2


Join at menti.com | use code 2224 1629
How much data is estimated to be generated every day worldwide?
None of the options are correct!
1 petabyte
100 terabytes
2.5 quintillion bytes


TM
Menti
BDTT-S1-2
Choose a slide to present


explainity
... explains
Big Data

Characteristics of Big Data (The 5 Vs)



Big Data vs Traditional Data

Aspect	Big Data	Traditional Data
Size	Large (terabytes/petabytes)	Smaller (megabytes/gigabytes)
Types	Structured, semi structured, unstructured	Primarily structured
Processing	Distributed systems (Hadoop, Spark)	Centralised Systems (SQL, RDBMS)
Speed	High velocity	Slower, batch processing
Storage	Scalable (HDFS, Cloud solutions)	Centralised, limited scalability
Analysis Tools	Advanced tools (ML, NoSQL, Spark)	Traditional tools (SQL, Excel)
Purpose	Strategic insights and innovation	Routine operations

Importance of Big Data

HEALTHCARE

RETAIL

FINANCE

MANUFACTURING

TRANSPORTATION

EDUCATION

Data Structures

Aspect	Structured	Unstructured	Semi-structured
Organisation	Fully organised (schema)	No predefined structure	Partially organised
Ease of analysis	Easy	Challenging	Moderate
Examples	SQL databases, financial reports	Social media, emails, videos	XML, JSON, NoSQL databases

```diff
+----------+-----+----------------+--------+
| Name     | Age | Address        | Salary |
+----------+-----+----------------+--------+
| John Doe | 30  | 123 Main St    | 50000  |
| Jane Doe | 28  | 456 Oak Lane   | 60000  |
+----------+-----+----------------+--------+
```

Data Structures

Aspect	Structured	Unstructured	Semi-structured
Organisation	Fully organised (schema)	No predefined structure	Partially organised
Ease of analysis	Easy	Challenging	Moderate
Examples	SQL databases, financial reports	Social media, emails, videos	XML, JSON, NoSQL databases

```json
{
  "Name": "John Doe",
  "Age": 30,
  "Address": {
    "Street": "123 Main St",
    "City": "London"
  },
  "Salary": 50000
}
```

```xml
<Employee>
  <Name>John Doe</Name>
  <Age>30</Age>
  <Address>
    <Street>123 Main St</Street>
    <City>London</City>
  </Address>
  <Salary>50000</Salary>
</Employee>
```

```arduino
Name,Age,Address
John Doe,30,"123 Main St, London"
Jane Doe,28,"456 Oak Lane, Manchester"
```

Data Structures

Aspect	Structured	Unstructured	Semi-structured
Organisation	Fully organised (schema)	No predefined structure	Partially organised
Ease of analysis	Easy	Challenging	Moderate
Examples	SQL databases, financial reports	Social media, emails, videos	XML, JSON, NoSQL databases

Data Sources for Big Data

- Social Media
- Internet of Things (IoT) Devices
- Transactional Data
- Machine-Generated Data
- Health Data
- Public Data
- Media and Entertainment
- Communication Platforms
- Satellite and Geospatial Data
- Cloud Services

Big Data Lifecycle

Data generation

Data collection

Data storage

Data processing and analysis

Data visualization and decision-making

Big Data Challenges

- Data Volume and Storage
- Data Integration
- Data Quality and Veracity
- Data Security and Privacy
- Talent Gap
- Data Governance

Meta’s AI Project Faces Privacy Complaints in Europe

The EU Is Taking on Big Tech. It May Be Outmatched

Emerging Job Roles in Big Data

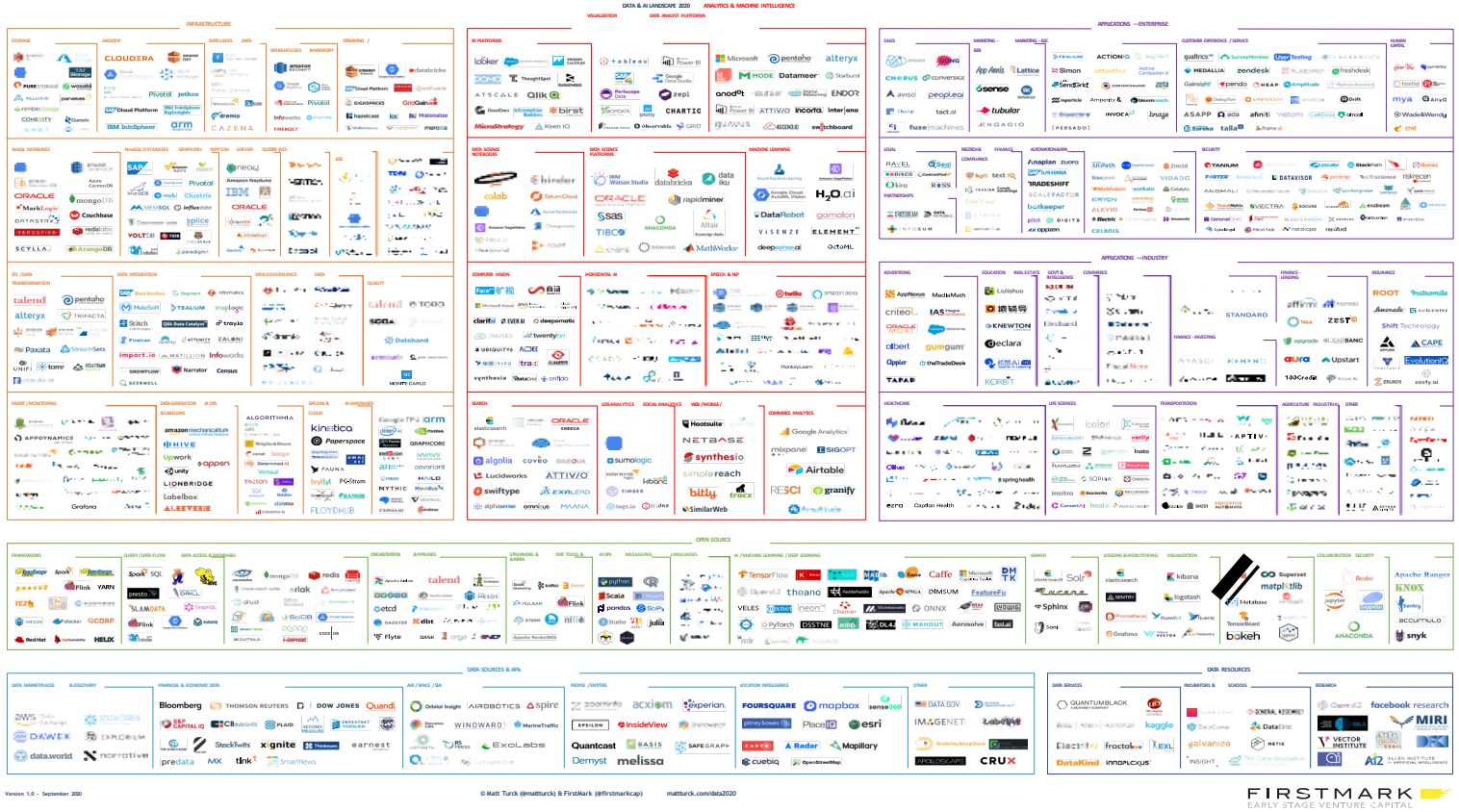
- Data Engineer
- Data Scientist
- Big Data Analyst
- Programming
- Machine Learning

How Facebook Tracks Your Data | NYT

Use Case

Netflix Recommendation System

Big Data Tools



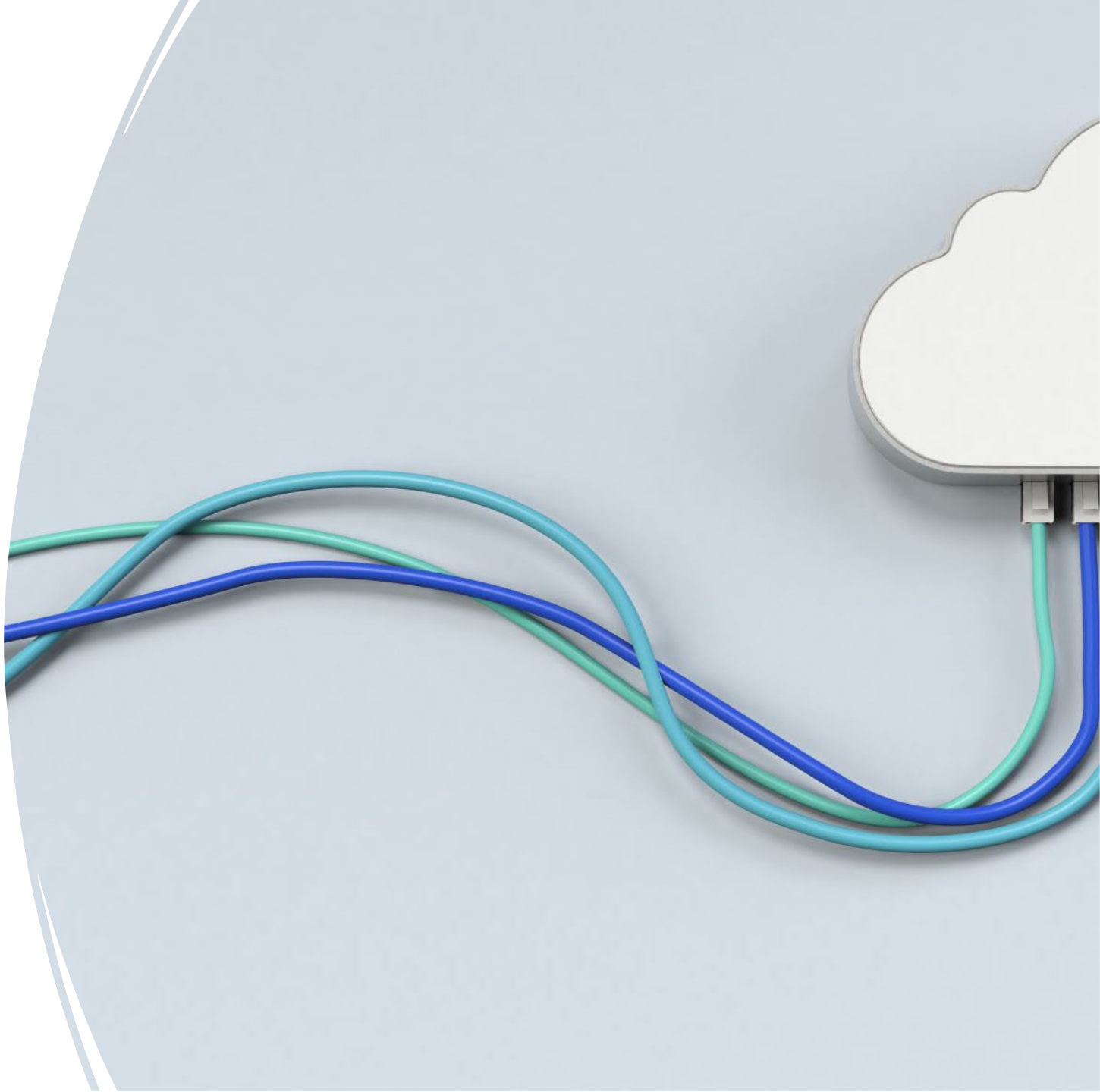
Cloud Computing

Cloud Computing?

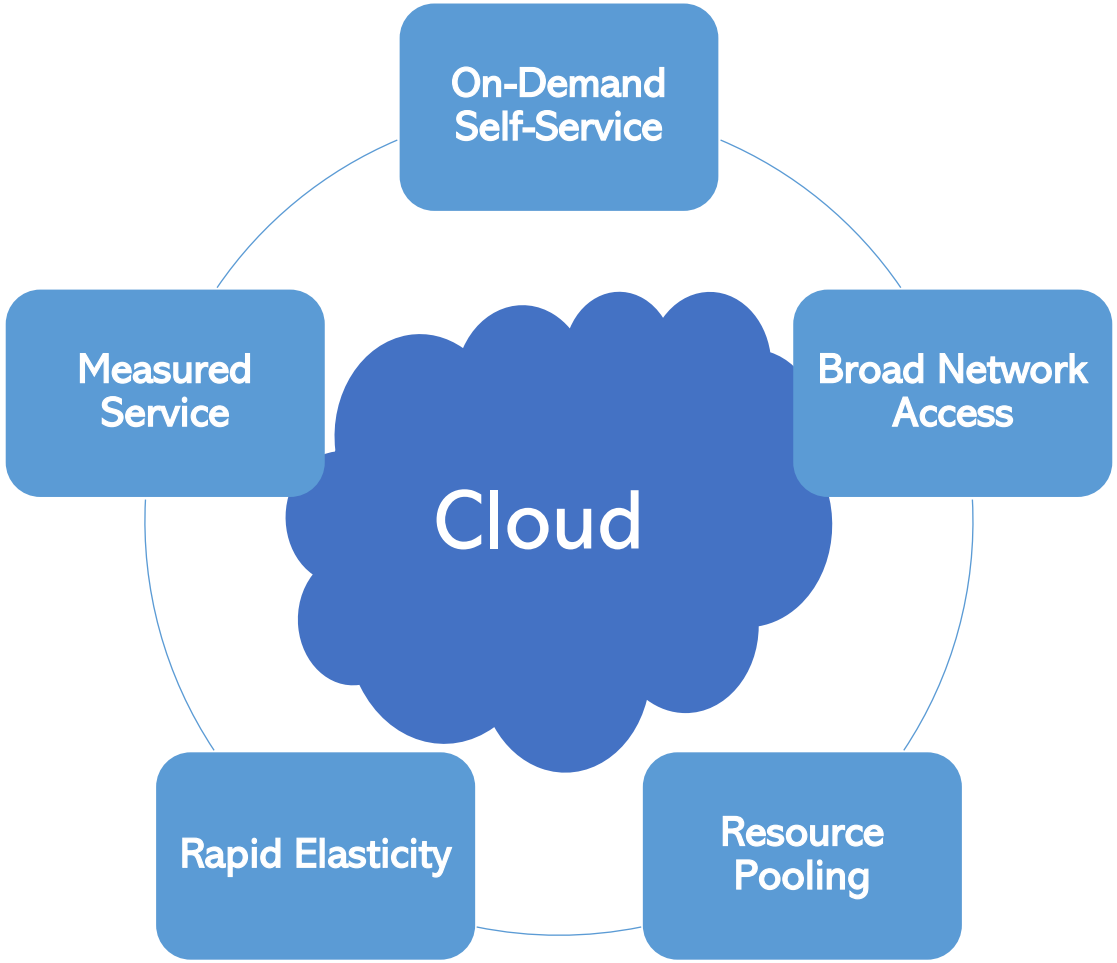
Cloud computing is best described as “a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources […] that can be rapidly provisioned and released with minimal management effort or service provider interaction” [NIST].

Cloud Computing in Big Data

- Scalability and Flexibility
- Cost-Effectiveness
- Enhanced Data Processing and Analytics
- Improved Collaboration and Accessibility
- Security and Compliance

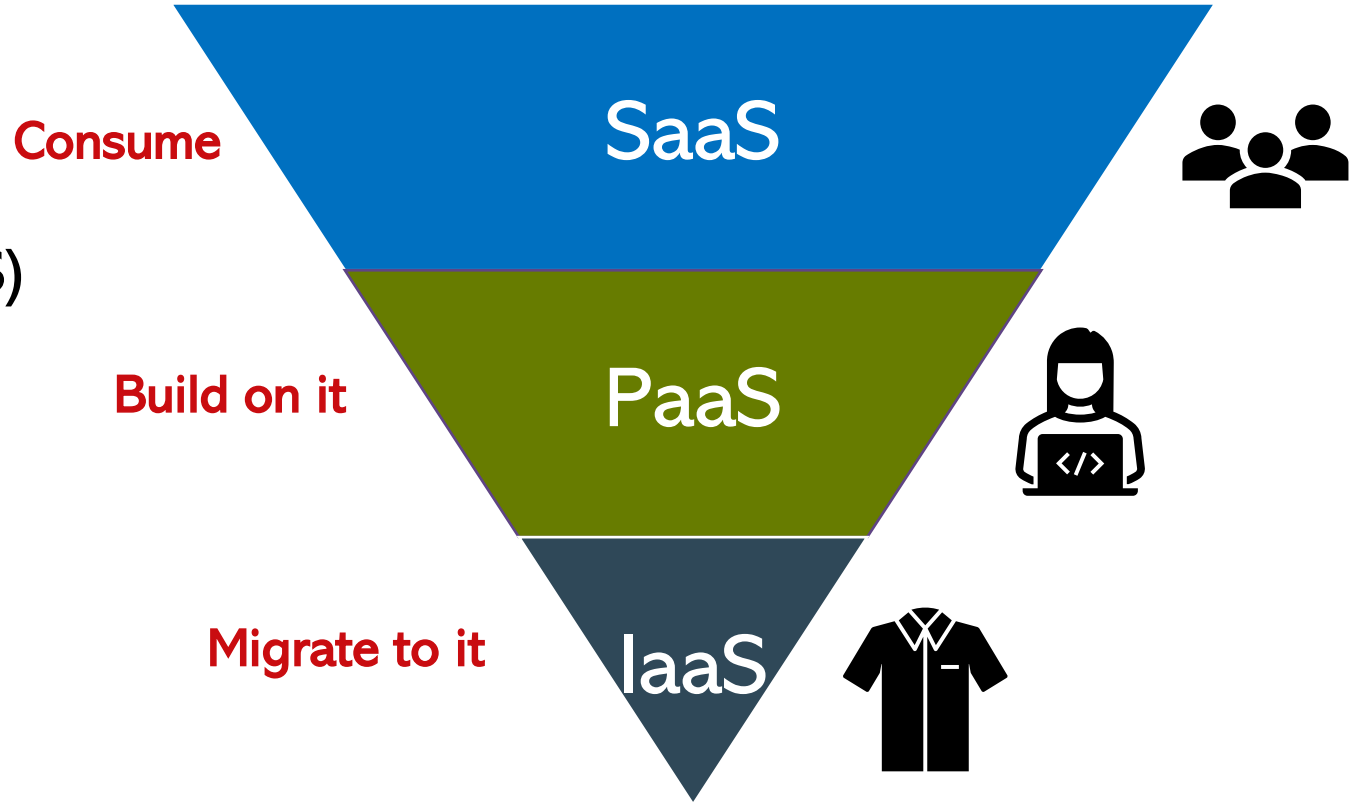


Key Characteristics of Cloud Computing



Cloud Service Models

- There are mainly 3 service models given as:
  - Software as a Service (SaaS)
  - Platform as a Service (PaaS)
  - Infrastructure as a Service (IaaS)



Deployment Models

Hybrid Cloud

- Bridges one or more Private, Public or Community clouds
- Allows manipulation of CapEx and OpEx to reduce costs
- Supports Resource Portability

Private Cloud

- Leverages existing CapEx
- Can help reduce OpEx
- Intended for a Single Tenant

Public Cloud

- Shifts CapEx to OpEx
- Offers a Pay as you go model
- Supports Multiple Tenants

Community Cloud

- Allows sharing of CapEx and OpEx to reduce costs
- Brings together groups or organisations with a common interest
- Supports Resource Portability

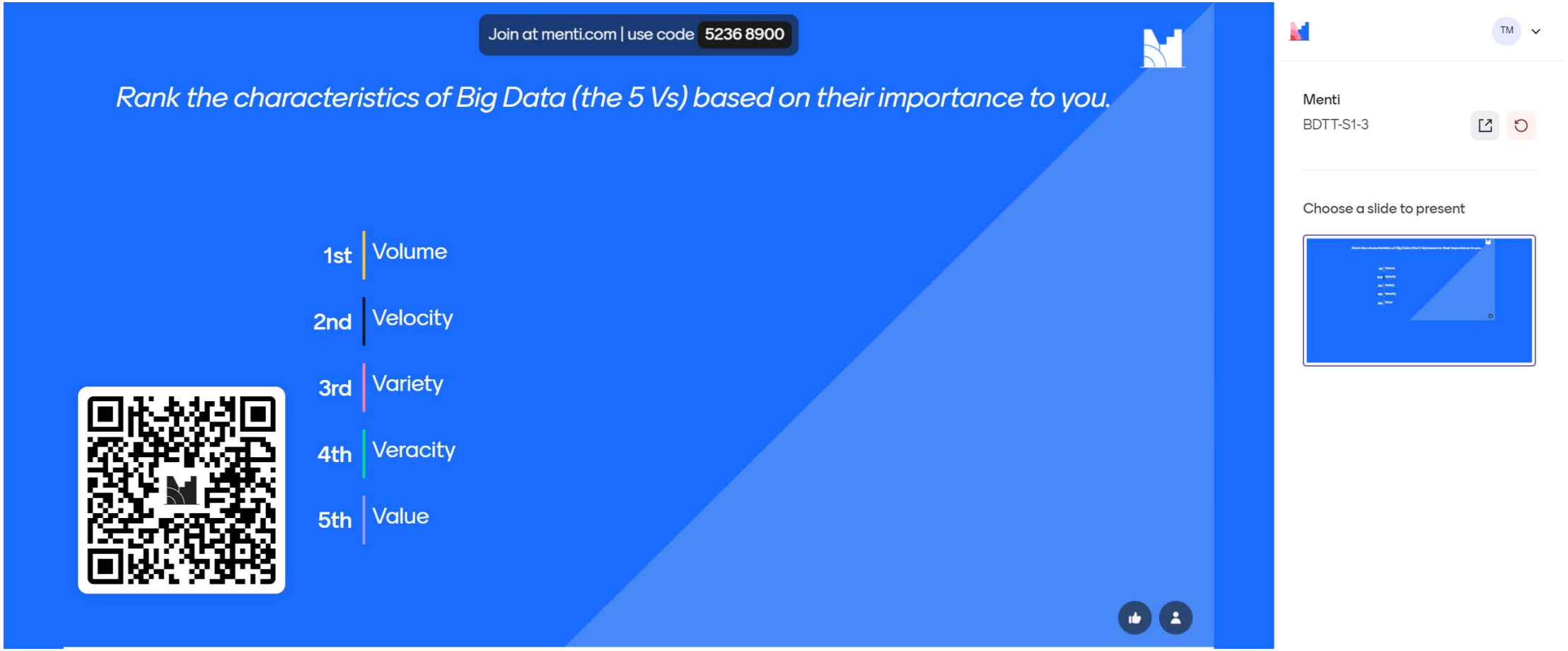
Tools that give you power

- Linux
- Python
- Spark (Databricks)
- MongoDB

Module Objectives

- Big data concepts and applications
- Appreciate the advantages of Cloud Computing
- How to process distributed data with Spark
  - Spark Core
  - Spark Streaming
  - Spark Machine Learning
  - Spark SQL
- How to manage data in NoSQL databases with MongoDB

Activity 3


Join at menti.com | use code 5236 8900
Rank the characteristics of Big Data (the 5 Vs) based on their importance to you.
1st Volume
2nd Velocity
3rd Variety
4th Veracity
5th Value
TM
Menti
BDTT-S1-3
Choose a slide to present

References

- https://www.geeksforgeeks.org/dikw-pyramid-data-information-knowledge-and-wisdom-data-science-and-big-data-analytics
- https://www.datacamp.com/cheat-sheet/the-data-information-knowledge-wisdom-pyramid
- https://iterationinsights.com/article/understanding-the-different-types-of-analytics/
- https://www.oracle.com/big-data/what-is-big-data/
- https://www.purestorage.com/knowledge/big-data/big-data-vs-traditional-data.html
- https://www.analyticsinsight.net/big-data-2/how-big-data-is-transforming-decision-making-across-industries
- https://www.ibm.com/think/topics/structured-vs-unstructured-data
- https://www.datamation.com/big-data/data-lifecycle-phases/
- https://www.forbes.com/advisor/business/what-is-cloud-computing/

Workshop

- Running a Databricks notebook
- Running Some Linux commands
- Working with Local Filesystem