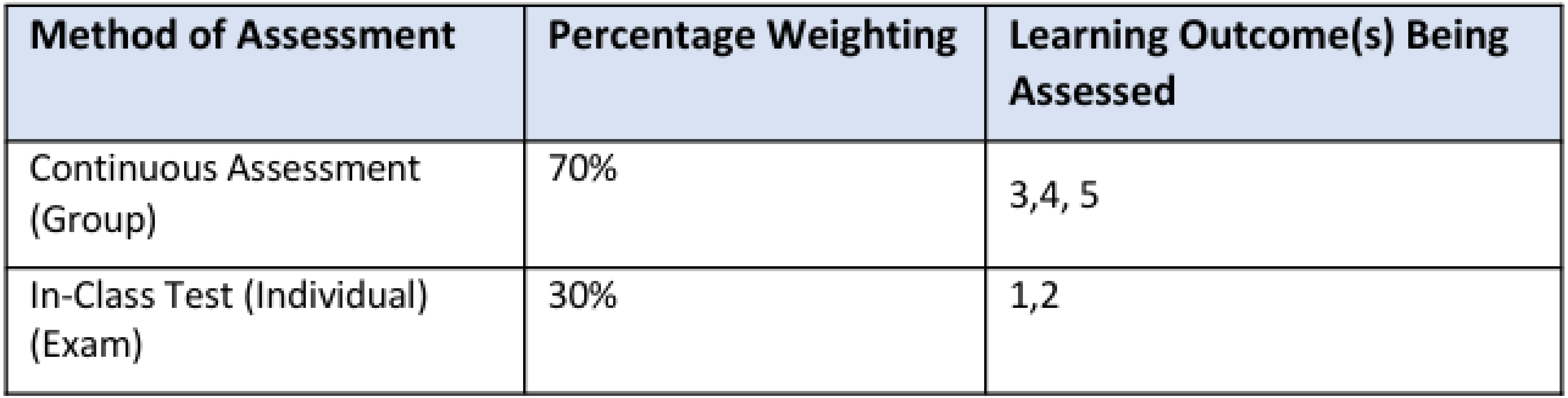
Lecture 1: Introduction to Big Data Systems

Dr Anesu Nyabadza

PhD: Engineering (Dublin City University)

Masters: Data Analytics and Professional development (The University of Sheffield)

Bachelors: Mechatronic Engineering (Dublin City University)



Method of Assessment	Percentage Weighting	Learning Outcome(s) Being Assessed
Continuous Assessment (Group)	70%	3,4, 5
In-Class Test (Individual) (Exam)	30%	1,2

Introduction to data storage for data analysis

Data storage plays a crucial role in the field of data analytics

Why do we need data storage for data analysis?

Volume and Variety→ Data analytics processes massive amounts of data from various sources. **Efficient data storage** is critical for managing this diversity and volume, as it allows for structured organization and easy retrieval.

Speed and Accessibility→ Quick access to data is critical for good analytics. Storage technology such as **database systems** provides real-time analysis and informed decision-making, which is critical for preserving a competitive advantage.

Scalability→ As an organization's data expands, scalable storage solutions are required to accommodate increased volumes without sacrificing speed and avoiding repeated system upgrades.

Why do we need data storage for data analysis?

Data Integrity and Reliability→Reliable storage ensures data accuracy and accessibility even during failures by utilizing redundancy, backups, and disaster recovery plans. This **ensures reliable analytics** and decision-making.

Security→ Storage systems must protect data from unwanted access and threats using mechanisms such as encryption and access controls to ensure that data stays confidential and integral.

Cost Efficiency→ Storage costs can be reduced by deduplicating data, compressing it, and selecting appropriate storage alternatives (cloud, on-premises, hybrid) depending on consumption patterns.

Regulatory Compliance→ Proper storage aids compliance with data management requirements by assuring **lawful handling, retention, and data sovereignty**.

Data has become as valuable as money in the business world

Informed Decision-Making→ Data enables organizations to make decisions based on facts, not intuition. This can result in more accurate and efficient strategic planning and resource allocation.

Enhanced Customer Insights→ Collecting and analyzing customer data assists businesses in understanding consumer behavior, preferences, and trends, allowing them to adjust their products and services to better suit customer needs.

Operational Efficiency→ Data analytics may discover operational inefficiencies and bottlenecks, allowing businesses to optimize processes, cut costs, and increase overall efficiency.

Data has become as valuable as money in the business world

Competitive Advantage→ Companies that efficiently use data can obtain a competitive advantage by forecasting market trends, optimizing marketing techniques, and inventing faster than competitors.

Risk management→ relies on data to forecast prospective risks and build mitigation solutions. This encompasses financial, market, and operational risks.

Personalization→ Data allows organizations to provide tailored experiences to their customers, which can boost engagement, contentment, and loyalty.

industry Trend Analysis→Analyzing data over time enables firms to detect and profit on industry trends and customer behavior patterns.

What Is A Database

What Is A Database

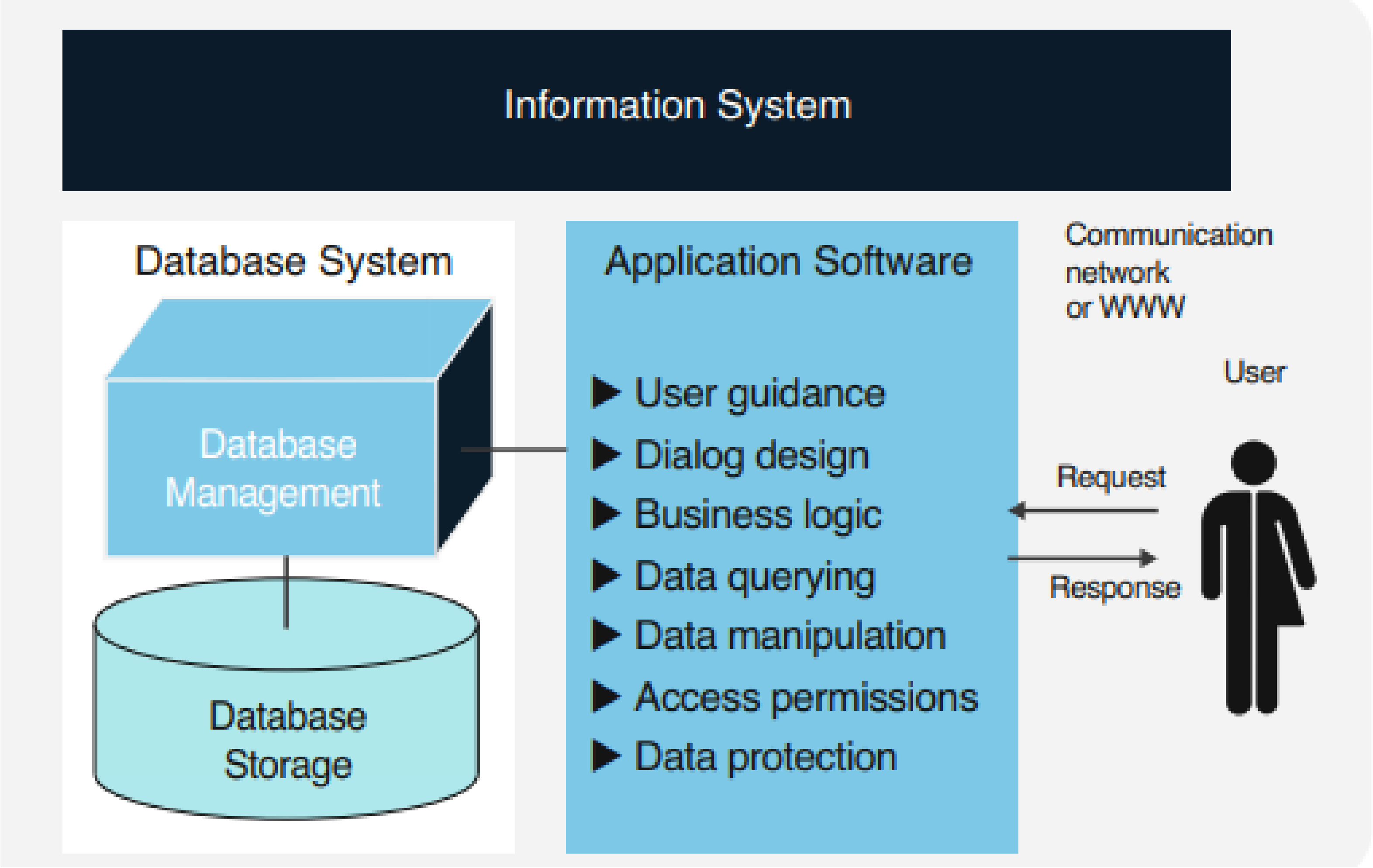
- A database is a technology for storing, retrieving and querying data.
- Data is structured into groups of similar data which allows data to be easily accessed and analysed.
- A database will have some kind of interface and/or language which allows data operations and queries.
- A database can run on a PC for one user or on a server for many users and systems to access
- The data may be structured in different ways depending on the type of database. This will change the way data is stored, manipulated and queried. In this course we will use two types of databases:
- **Relational Database**
- **NoSQL Database**

What Is A Database

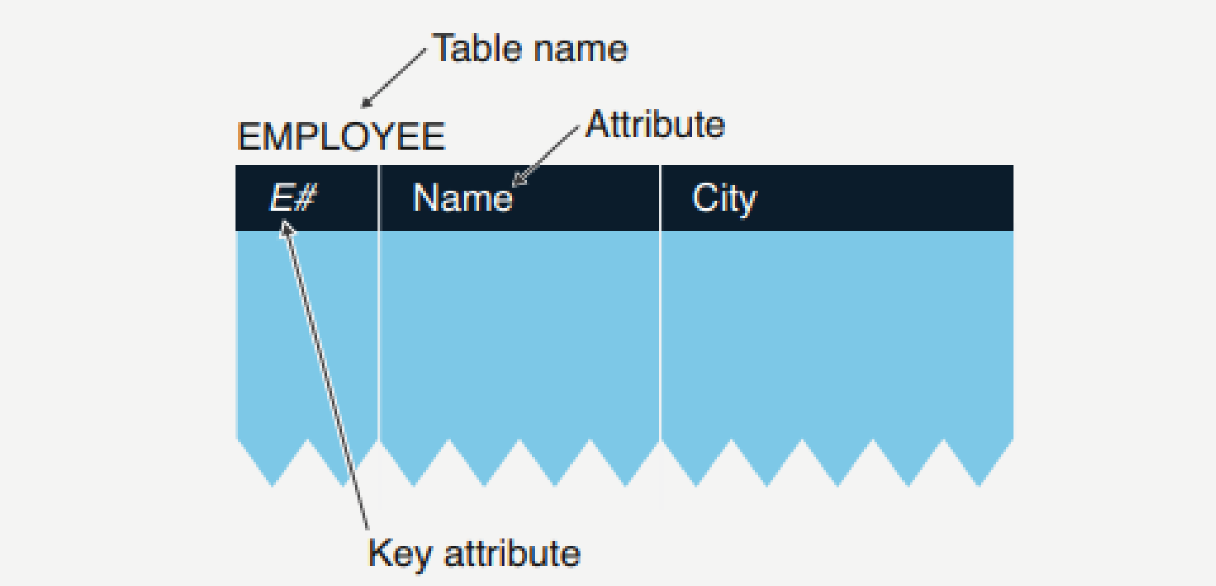
The evolution from the industrial society via the service society to the information and knowledge society is represented by the assessment of information as a factor in production. The following characteristics distinguish information from material goods:

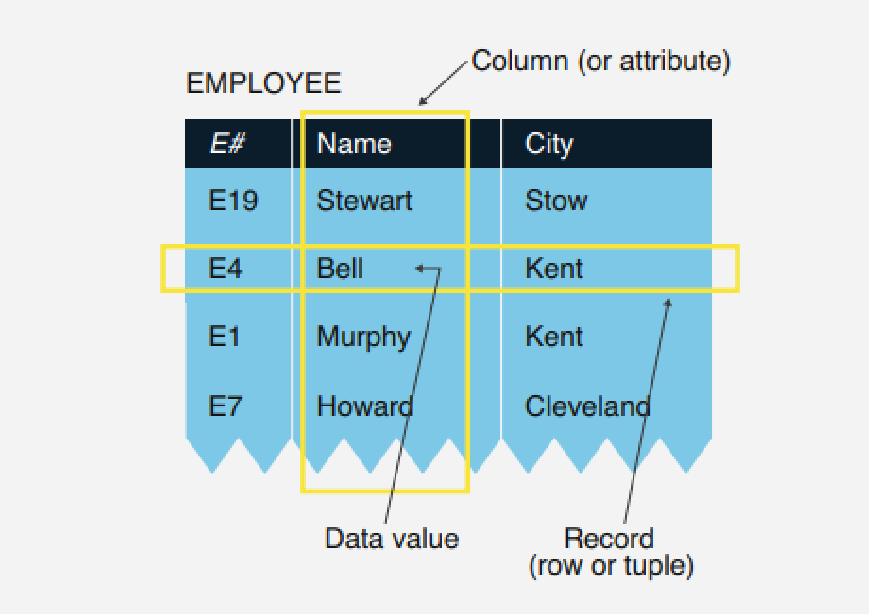
- Representation: Information is specified by data (signs, signals, messages, or language elements).
- Processing: Information can be transmitted, stored, categorized, found, or converted into other representation formats using algorithms and data structures (calculation rules).
- Combination: Information can be freely combined. The origin of individual parts cannot be traced. Manipulation is possible at any point.
- Age: Information is not subject to physical aging processes.
- Original: Information can be copied without limit and does not distinguish between original and copy.
- Vagueness: Information can be imprecise and of differing validity (quality).
- Medium: Information does not require a fixed medium and is therefore independent of location.

What Is A Database



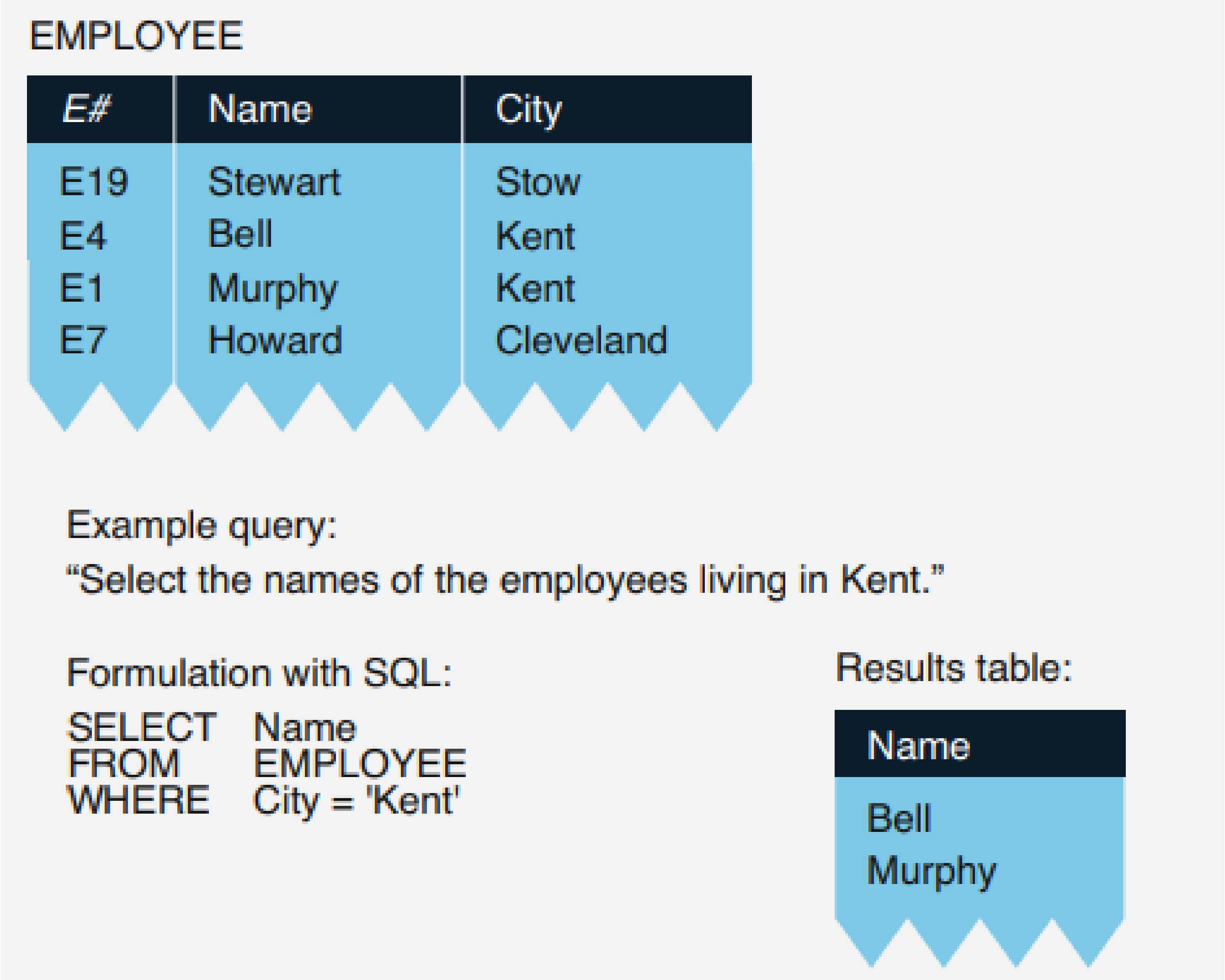
What Is A Database





E#	Name	City
E19	Stewart	Stow
E4	Bell	Kent
E1	Murphy	Kent
E7	Howard	Cleveland

SQL Syntax



EMPLOYEE

E#	Name	City
E19	Stewart	Stow
E4	Bell	Kent
E1	Murphy	Kent
E7	Howard	Cleveland

Example query:
“Select the names of the employees living in Kent.”

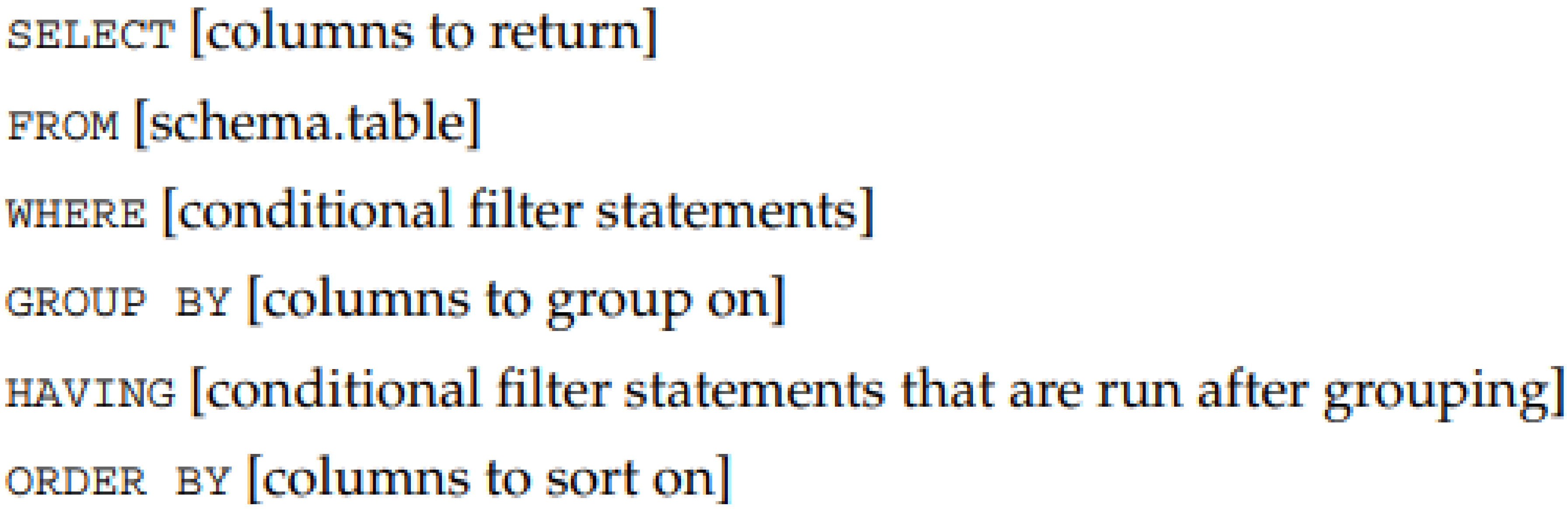
Formulation with SQL:

```
SELECT   Name
FROM     EMPLOYEE
WHERE    City = 'Kent'
```

Results table:

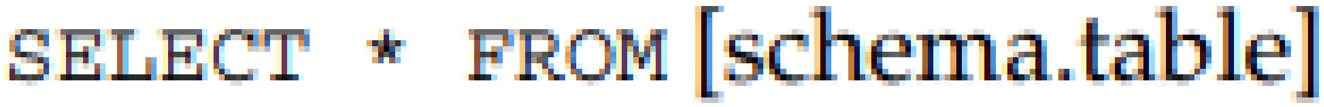
Name
Bell
Murphy

SQL Syntax



```
SELECT [columns to return]
FROM [schema.table]
WHERE [conditional filter statements]
GROUP BY [columns to group on]
HAVING [conditional filter statements that are run after grouping]
ORDER BY [columns to sort on]
```

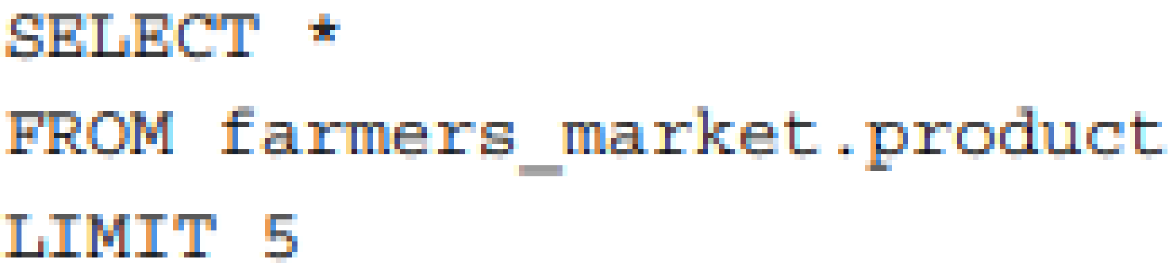
SQL Syntax



```
SELECT * FROM [schema.table]
```

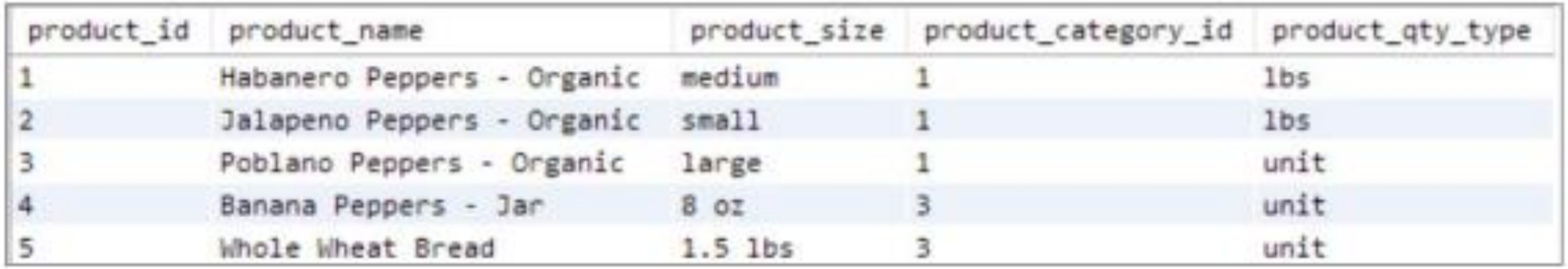
- This selects the entire table (*)
- SQL functions are not case-sensitive (i.e., you can use either SELECT or select)

SQL Syntax



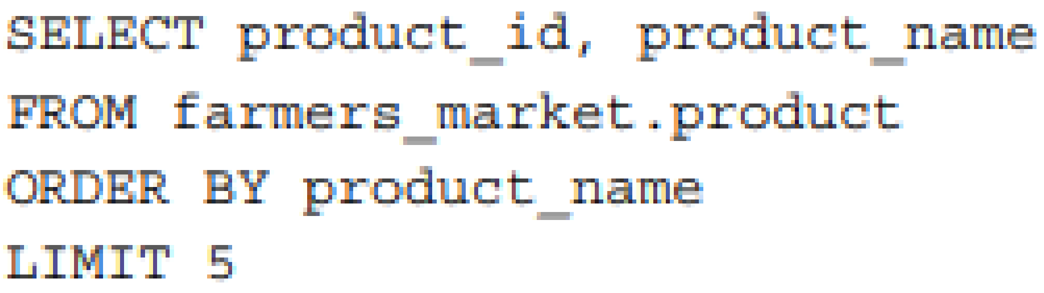
```
SELECT *
FROM farmers_market.product
LIMIT 5
```

- The above query returns all columns for the first 5 rows of the product table



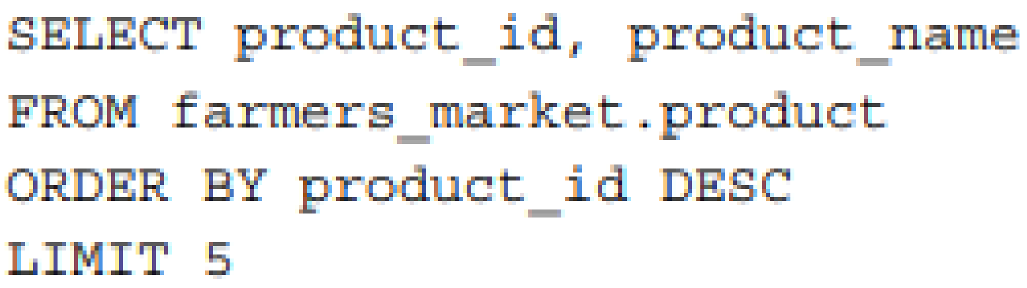
product_id	product_name	product_size	product_category_id	product_qty_type
1	Habanero Peppers - Organic	medium	1	lbs
2	Jalapeno Peppers - Organic	small	1	lbs
3	Poblano Peppers - Organic	large	1	unit
4	Banana Peppers - Jar	8 oz	3	unit
5	Whole Wheat Bread	1.5 lbs	3	unit

SQL Syntax



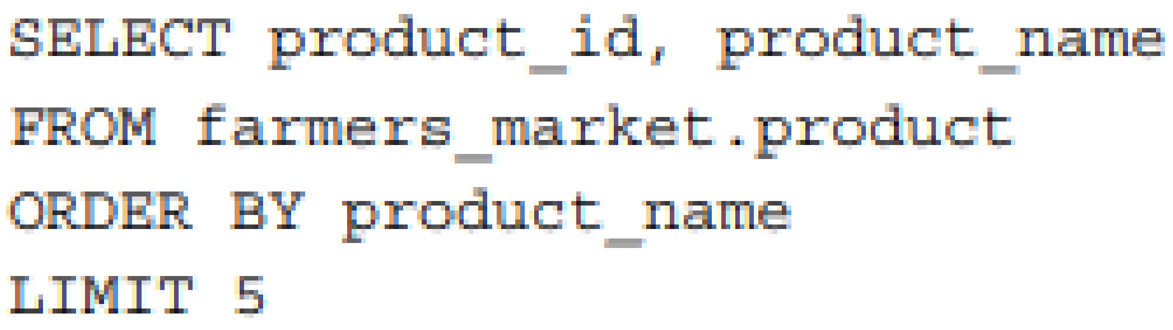
```
SELECT product_id, product_name
FROM farmers_market.product
ORDER BY product_name
LIMIT 5
```

- The above QUERY sorts the results by product name, even though the product ID is listed first in the output
- the following modification to the ORDER BY clause changes the query to now sort the results by product ID, highest to lowest



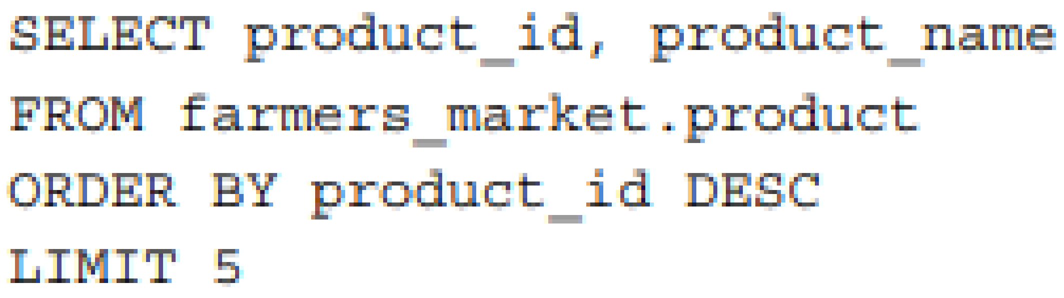
```
SELECT product_id, product_name
FROM farmers_market.product
ORDER BY product_id DESC
LIMIT 5
```

SQL Syntax



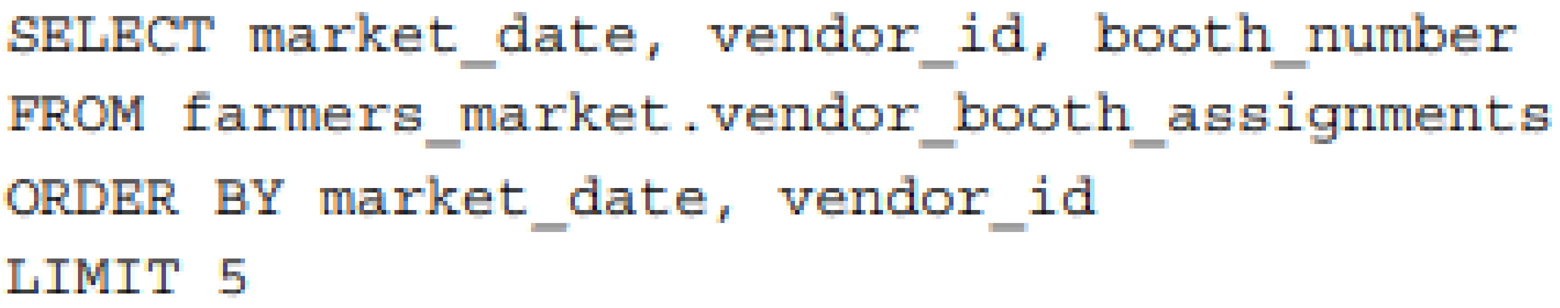
```
SELECT product_id, product_name
FROM farmers_market.product
ORDER BY product_name
LIMIT 5
```

- The above QUERY sorts the results by product name, even though the product ID is listed first in the output
- the following modification to the ORDER BY clause changes the query to now sort the results by product ID, highest to lowest

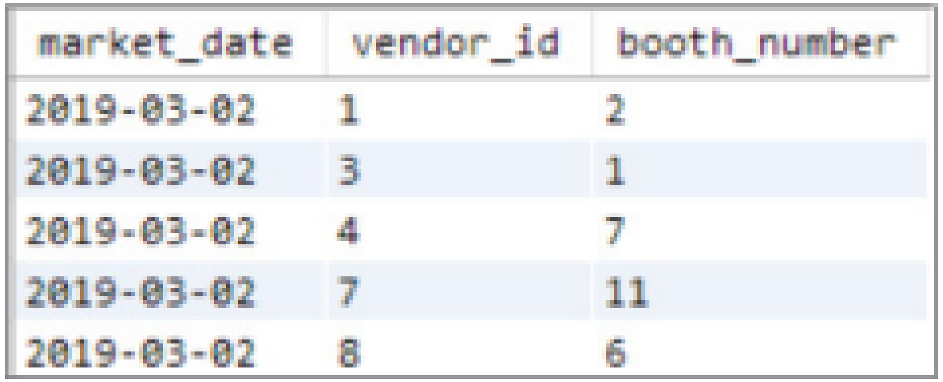


```
SELECT product_id, product_name
FROM farmers_market.product
ORDER BY product_id DESC
LIMIT 5
```

In the following modification after sorting by market date, the records are then sorted by vendor ID
in ascending order by default (USE ***ASC*** *TO SPECIFY ASCENDING ORDER*)



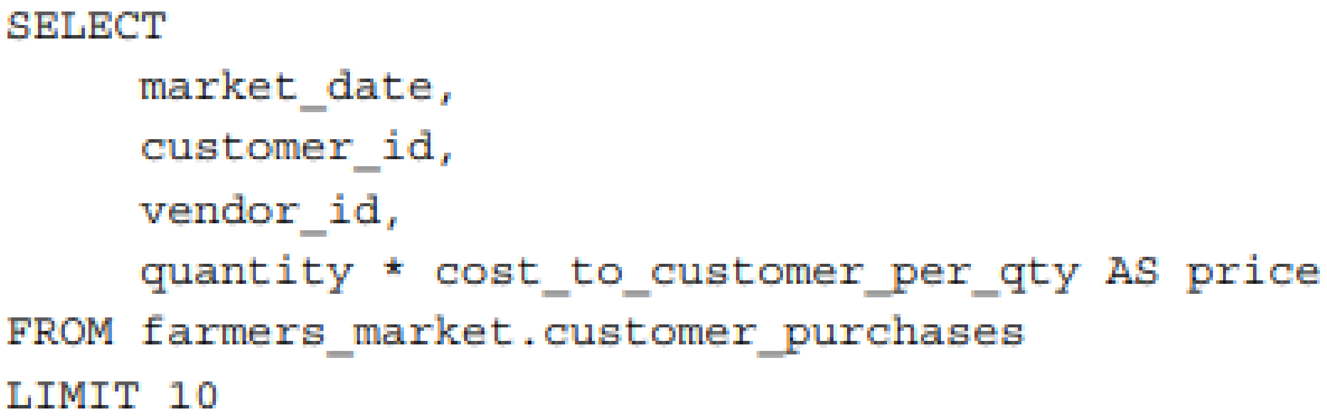
```
SELECT market_date, vendor_id, booth_number
FROM farmers_market.vendor_booth_assignments
ORDER BY market_date, vendor_id
LIMIT 5
```



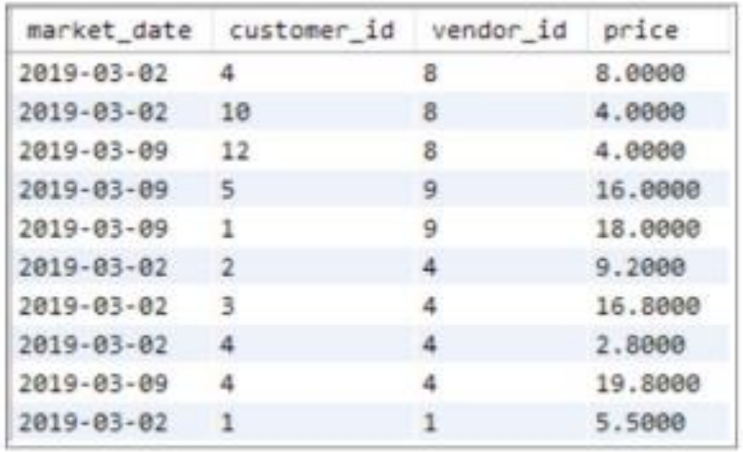
market_date	vendor_id	booth_number
2019-03-02	1	2
2019-03-02	3	1
2019-03-02	4	7
2019-03-02	7	11
2019-03-02	8	6

SQL Syntax

The * symbol is also used to multiply variables



```
SELECT
     market_date,
     customer_id,
     vendor_id,
     quantity * cost_to_customer_per_qty AS price
FROM farmers_market.customer_purchases
LIMIT 10
```



market_date	customer_id	vendor_id	price
2019-03-02	4	8	8.0000
2019-03-02	10	8	4.0000
2019-03-09	12	8	4.0000
2019-03-09	5	9	16.0000
2019-03-09	1	9	18.0000
2019-03-02	2	4	9.2000
2019-03-02	3	4	16.8000
2019-03-02	4	4	2.8000
2019-03-09	4	4	19.8000
2019-03-02	1	1	5.5000

Importance of databases

Structured Data Governance: SQL databases excel in managing structured data, a crucial aspect in big data scenarios. While big data involves diverse data types, the structured nature of SQL databases proves beneficial for handling certain elements like metadata, configuration information, or specific transactional data.

Integration with Analytics Tools: Business intelligence and analytics platforms seamlessly integrate with SQL databases. Leveraging SQL databases in big data environments allows for smooth integration with these tools, enabling organizations to extract meaningful insights from their data.

Data Consistency and Integrity Assurance: SQL databases enforce data consistency and integrity through features such as transactions, constraints, and foreign keys. In big data applications where data accuracy is paramount, SQL databases provide a robust framework for ensuring data consistency.

Scalability and Performance Enhancement: SQL databases provide features for optimizing performance and scalability, including indexing, query optimization, and caching mechanisms. In big data technologies involving distributed and NoSQL databases, SQL databases contribute to specific components of a big data architecture, optimizing performance.

Importance of databases

Interoperability Standardization→ SQL, as a widely adopted standard, ensures interoperability across various systems and tools. This is particularly valuable in big data ecosystems where diverse technologies coexist, allowing seamless integration.

Security Measures→ SQL databases boast strong security features, including authentication, authorization, and access control. In big data environments handling sensitive information, robust security measures are crucial, and SQL databases contribute to meeting these security requirements.

Historical Data Management→SQL databases are well-suited for efficiently managing historical data, enabling organizations to store and retrieve records. In big data scenarios, where historical data analysis is vital for trend analysis and decision-making, SQL databases complement other data storage solutions.

Data Warehousing Functions→ SQL databases commonly support data warehousing, involving the collection, storage, and management of large data volumes for reporting and analysis. In big data architectures, SQL-based data warehousing solutions can be integrated to support analytical processing and reporting needs.

Big Data

Data
Engineer

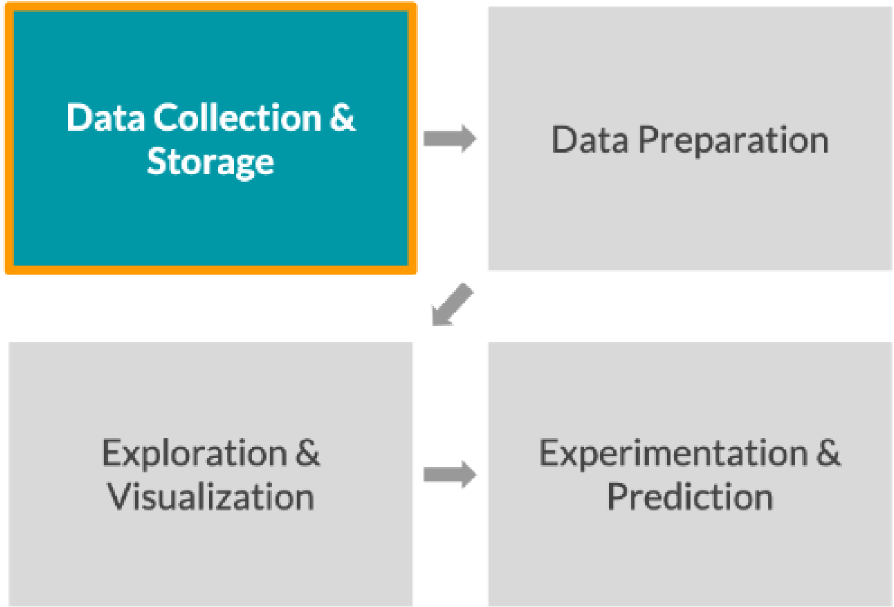
Data
Analyst

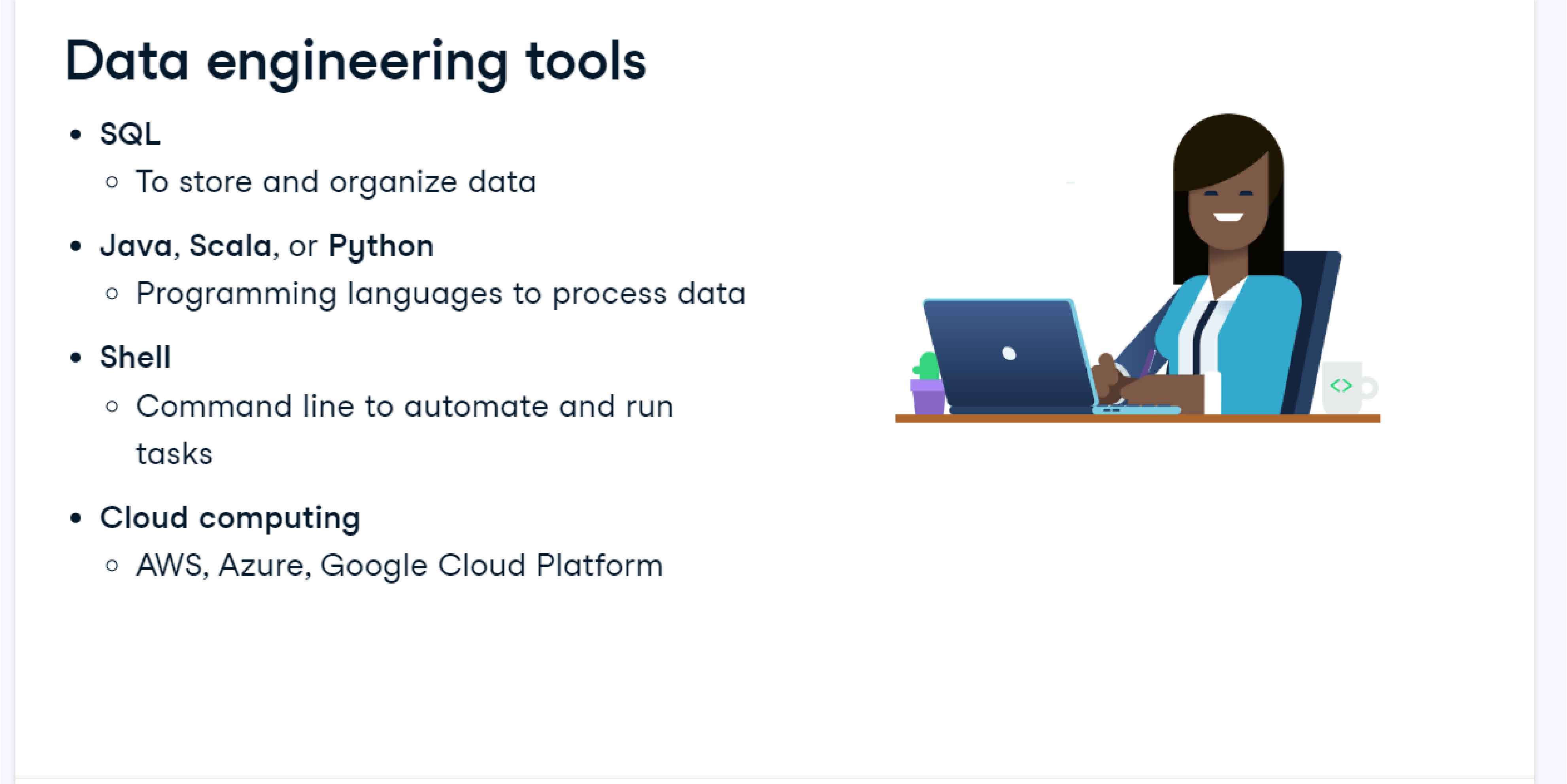
Data
Scientist

Machine Learning
Scientist

Data engineer

- Information architects
- Build data pipelines and storage solutions
- Maintain data access


Data Collection & Storage
Data Preparation
Exploration & Visualization
Experimentation & Prediction



Data engineering tools

- **SQL**
  - To store and organize data
- **Java**, **Scala**, or **Python**
  - Programming languages to process data
- **Shell**
  - Command line to automate and run tasks
- **Cloud computing**
  - AWS, Azure, Google Cloud Platform

Data analyst tools

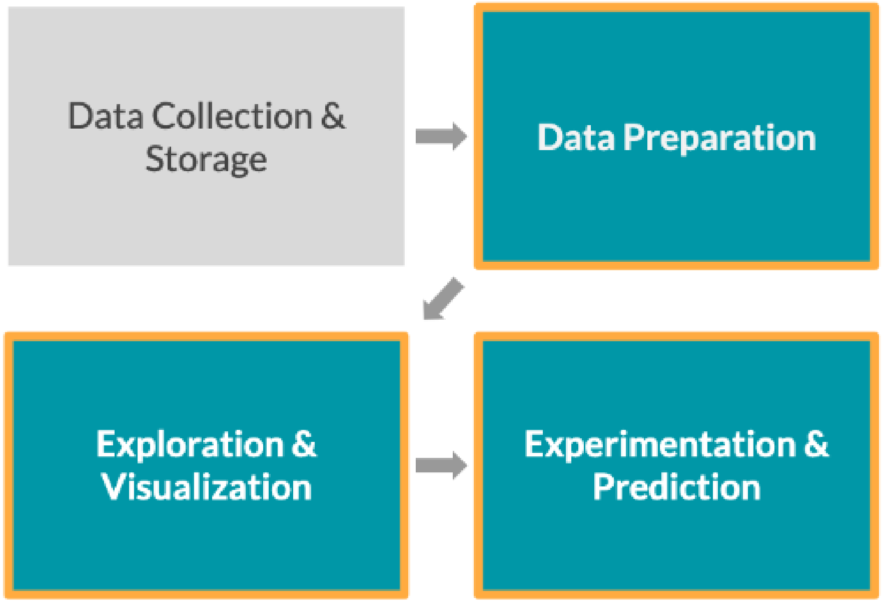
- **SQL**
  - Retrieve and aggregate data
- **Spreadsheets (Excel or Google Sheets)**
  - Simple analysis
- **BI tools (Tableau, Power BI, Looker)**
  - Dashboards and visualizations
- *May have:* Python or R
  - Clean and analyze data

Data analyst tools

- **SQL**
  - Retrieve and aggregate data
- **Spreadsheets (Excel or Google Sheets)**
  - Simple analysis
- **BI tools (Tableau, Power BI, Looker)**
  - Dashboards and visualizations
- *May have:* Python or R
  - Clean and analyze data

Data scientist

- Versed in statistical methods
- Run experiments and analyses for insights
- Traditional machine learning



Data Collection & Storage

Data Preparation

Exploration & Visualization

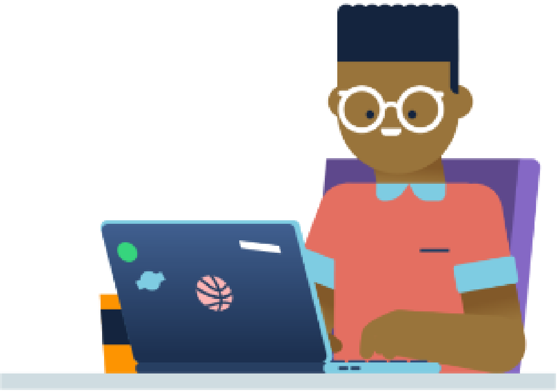
Experimentation & Prediction

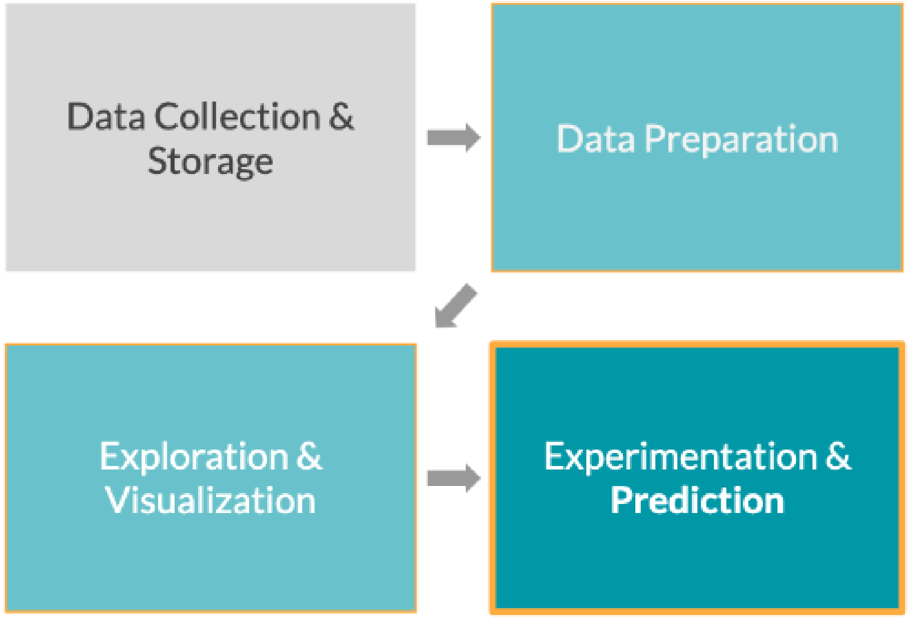
Data scientist tools

- **SQL**
  - Retrieve and aggregate data
- **Python and/or R**
  - Data science libraries, e.g., `pandas` (Python) and `tidyverse` (R)

Machine learning scientist

- Predictions and extrapolations
- Classification
- Deep learning
  - Image processing
  - Natural language processing





Data Collection & Storage

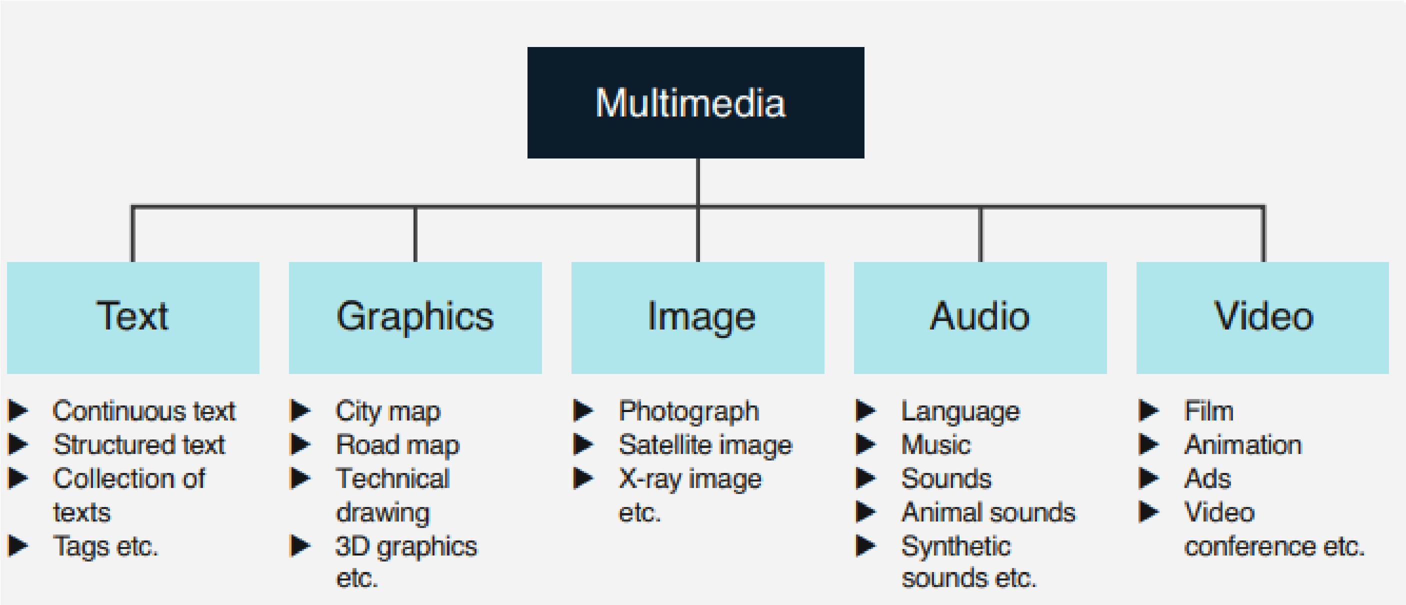
Data Preparation

Exploration & Visualization

Experimentation & **Prediction**

Big Data

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”



Multimedia

Text
- Continuous text
- Structured text
- Collection of texts
- Tags etc.

Graphics
- City map
- Road map
- Technical drawing
- 3D graphics etc.

Image
- Photograph
- Satellite image
- X-ray image etc.

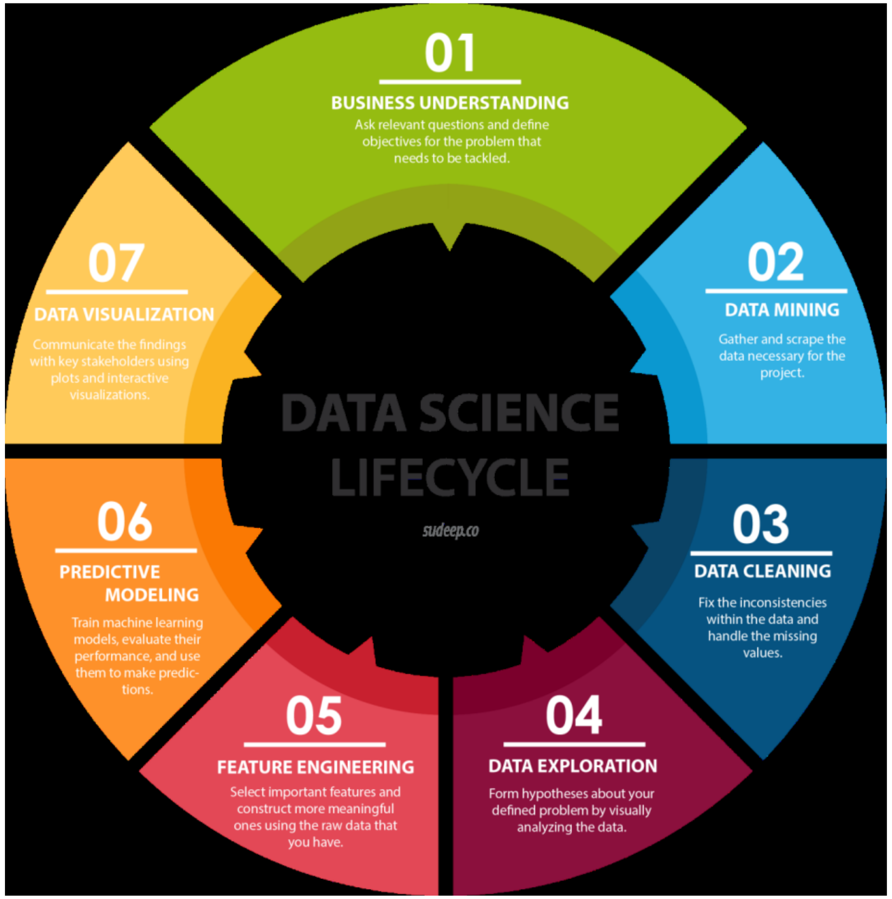
Audio
- Language
- Music
- Sounds
- Animal sounds
- Synthetic sounds etc.

Video
- Film
- Animation
- Ads
- Video conference etc.

Big Data

All use of Big Data applications requires successful management of the three Vs
Volume: There are massive amounts of data involved, ranging from giga- to zettabytes (megabyte, 106 bytes; gigabyte, 109 bytes; terabyte, 1012 bytes; petabyte, 1015 bytes; exabyte, 1018 bytes; zettabyte, 1021 bytes).
Variety: Big Data involves storing structured, semi-structured, and unstructured multimedia data (text, graphics, images, audio, and video; cf. Fig. 1.7).
Velocity: Applications must be able to process and analyze data streams in real time as the data is gathered.

Big Data can be considered an information asset, which is why sometimes another V is added:
Value: Big Data applications are meant to increase the enterprise value, so investments in personnel and technical infrastructure are made where they will bring leverage or added value can be generated.
To complete our discussion of the concept of Big Data, we will look at another V:
Veracity: Since much data is vague or inaccurate, specific algorithms evaluating the validity and assessing result quality are needed. Large amounts of data do not automatically mean better analyses.
Veracity is an important factor in Big Data, where the available data is of variable quality, which must be taken into consideration in analyses. Aside from statistical methods, there are fuzzy methods of soft computing that assign a truth value between 0 (false) and 1 (true) to any result or statement.



DATA SCIENCE LIFECYCLE

01 BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING
Gather and scrape the data necessary for the project.

03 DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

sudeep.co

Introduction To Databases & Queries Setting Up the Program.

- This lesson will introduce you to relational databases, how to set them up, import data into them and to perform basic queries on single tables.
- XAMPP and MySQL are the key technologies used in this module.
- XAMPP stands for Cross-Platform (X), Apache (A), MariaDB (M), PHP (P) and Perl (P). It is a simple, lightweight Apache distribution that makes it extremely easy for developers to create a local web server for testing and deployment purposes.
- MySQL is an open-source relational database management system. Its name contains "SQL", the abbreviation for Structured Query Language. Structured Query Language (SQL) is a standard computer language for relational database management and data manipulation. SQL is used to query, insert, update and modify data. In database programming, create, read, update, and delete(CRUD) are often referred to as the four basic functions of persistent storage.
- We will initiall focus on the Read or Query function to retrieve information from a database. This is the key activity of Data Scientists.
- In relational databases, and flat file databases, a table is a set of data elements (values) using a model of vertical columns (identifiable by name, and also called fields) and horizontal rows (also known as records), the cell is the unit where a row and column intersect. A table has a specified number of columns but can have any number of rows. Each row is identified by one or more values appearing in a particular column subset. A specific choice of columns that uniquely identify rows is called the primary key.

1. XAMPP & MySQL

- Open XAMPP, this can be found in the programs on your computer (this is also free to download) Possible link to use

  https://www.apachefriends.org/download.html

- XAMPP is a free open source package that allows us to use MySQL. MySQL is a database programming language.
- Click start on Apache
- Click start on MySQL.
- Click on Admin for MySQL to open the PHPAdmin browser-based interface which allows you to connect to and manipulate the database application.
- Once you have accessed phpMyAdmin, you first need to create a new database. Click on the “New” option in the navigation panel on the left. Enter an appropriate name for your database and click the “Create” button.