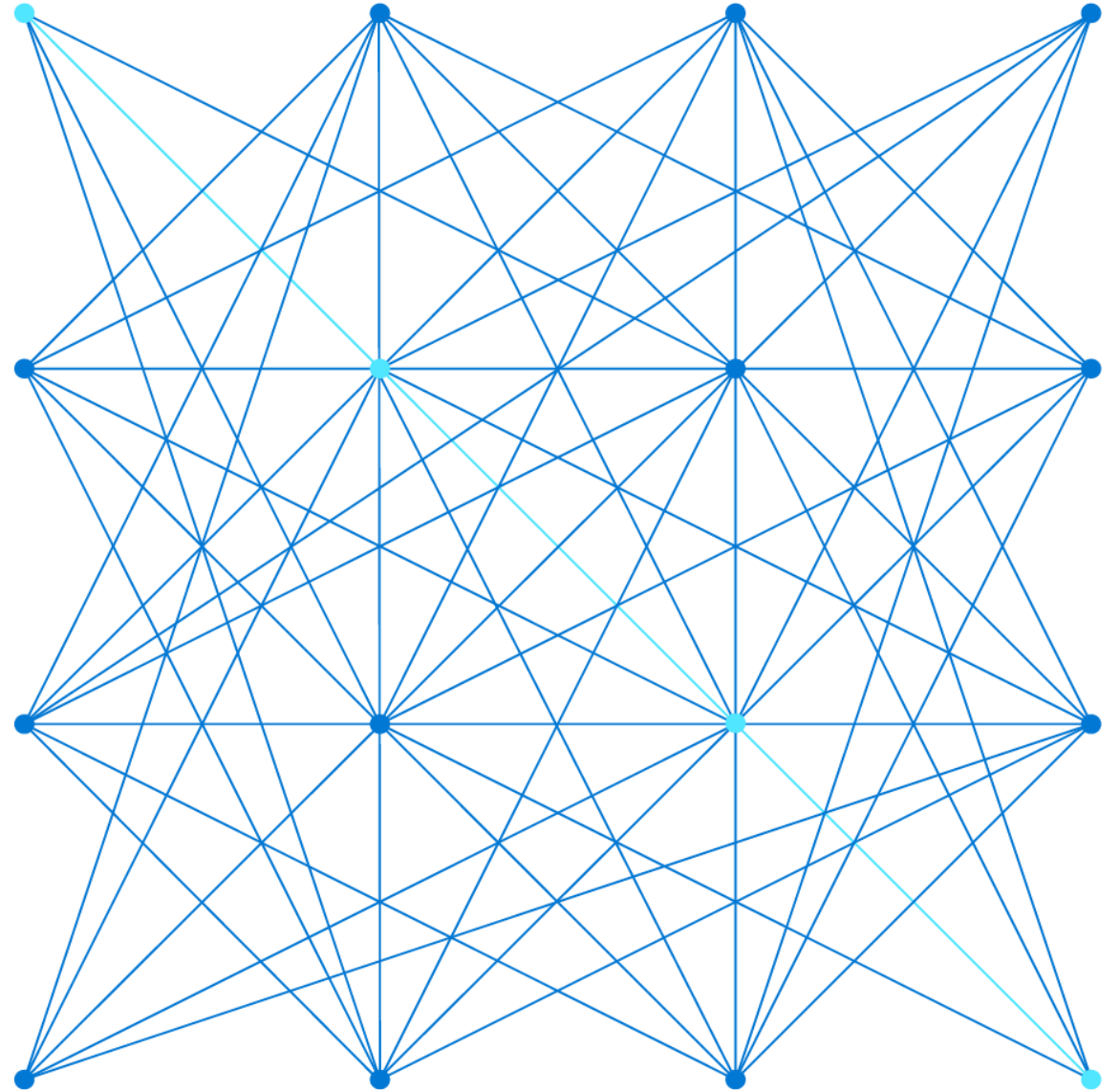


# Module 1: Explore core data concepts

Mohammed Arif  
10/03/2022



# Agenda



Explore core data concepts



Explore roles and responsibilities in the world of data (optional)



Describe concepts of relational data

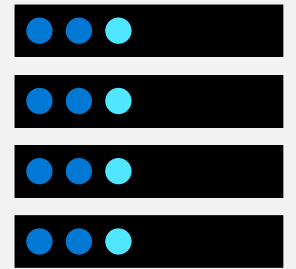


Explore concepts of non-relational data

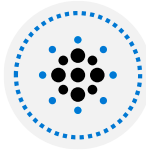


Explore concepts of data analytics

# Lesson 1: Explore core data concepts



# Lesson 1 objectives



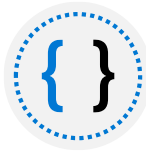
Identify how data is defined and stored



Identify characteristics of relational and non-relational data



Describe and differentiate data workloads



Describe and differentiate batch and streaming data

# What is data?

Collection of facts, numbers, descriptions, objects , stored in a structured, semi-structured, unstructured way

Structured

Table

|  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |
|--|--|--|
|  |  |  |
|  |  |  |

|  |  |  |  |  |  |
|--|--|--|--|--|--|
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |


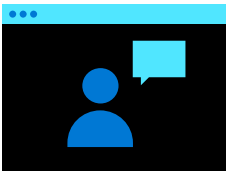
|  |  |  |
|--|--|--|
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

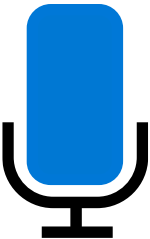

|  |  |  |
|--|--|--|
|  |  |  |
|  |  |  |

Semi-structured

```
## Document 1 ## {
"customerID": "103248",
"name": { "first": "AAA",
"last": "BBB" }, "address": {
"street": "Main Street",
"number": "101", "city":
"Acity", "state": "NY" },
"ccOnFile": "yes",
"firstOrder": "02/28/2003" }
## Document 2 ## {
"customerID": "103249",
"name": { "title": "Mr",
"forename": "AAA",
"lastname": "BBB" },
"address": { "street":
"Another Street", "number":
"202", "city": "Bcity",
"county": "Gloucestershire",
"country-region": "UK" },
"ccOnFile": "yes" }
```

Unstructured





# Transactional vs analytical data stores

## Online Transactional Processing (OLTP)

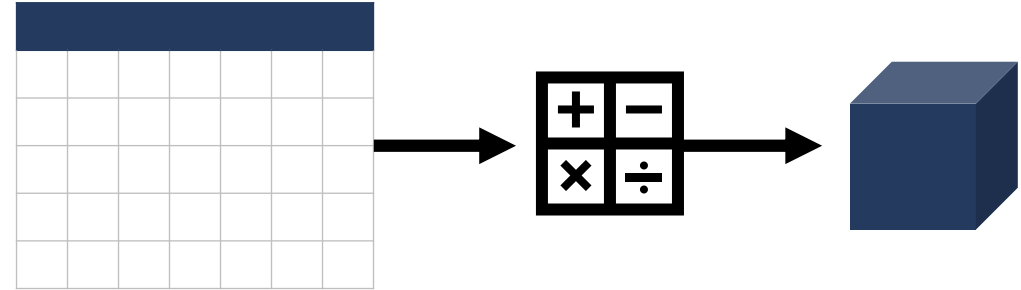
| Customer   |              |               |
|------------|--------------|---------------|
| CustomerID | CustomerName | CustomerPhone |
|            |              |               |
|            |              |               |

| Orders  |            |           |
|---------|------------|-----------|
| OrderID | CustomerID | OrderDate |
|         |            |           |
|         |            |           |

Data is stored one transaction at a time

## Online Analytical Processing (OLAP)

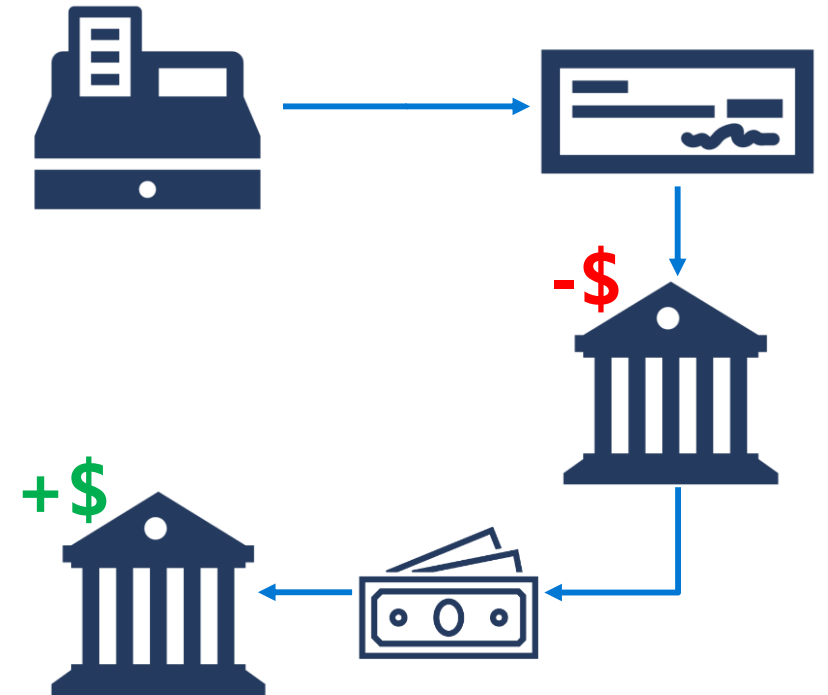


Data is periodically loaded, aggregated and stored in a cube

# Transactional workloads

Transactional data is information that tracks the interactions related to an organization's activities.

- **Atomicity** – each transaction is treated as a single unit, which success completely or fails completely.
- **Consistency** – transactions can only take the data in the database from one valid state to another.
- **Isolation** – concurrent execution of transactions leave the database in the same state.
- **Durability** – once a transaction has been committed, it will remain committed.



# Analytical Workloads

Analytical workloads are used for data analysis and decision making.

- Summaries
- Trends
- Business information

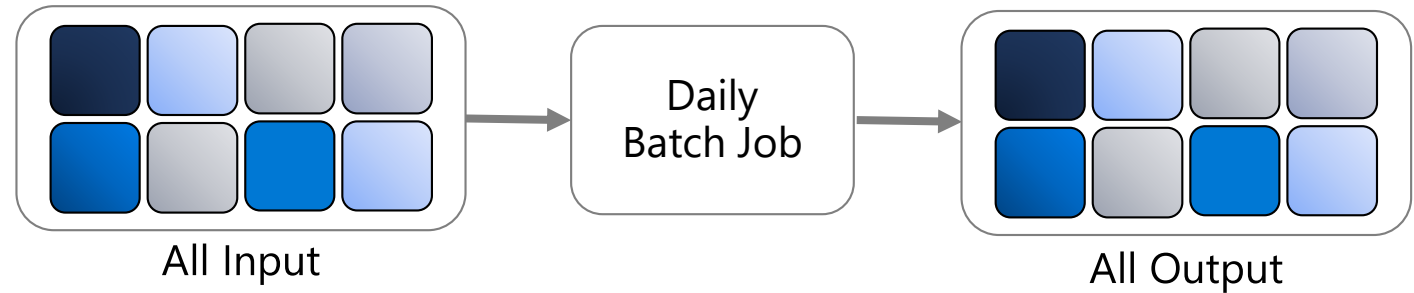




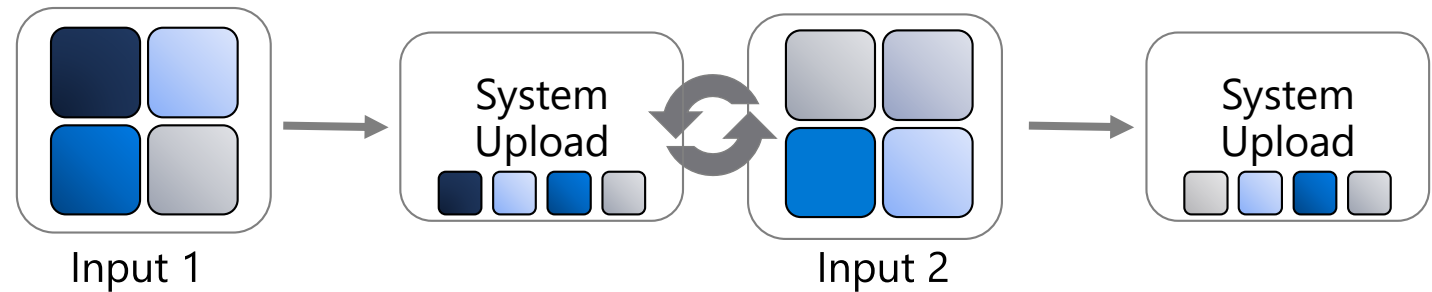
# Data Processing

Data processing is the conversion of raw data to meaningful information through a process.

**Batch Processing:** data elements are collected into a group. The whole group is then processed at a future time as a batch



**Stream Processing:** each new piece of data is processed when it arrives.



# Lesson 1: Knowledge check



**How is data in a relational table organized?**

- ☒ Rows and Columns
  - ☐ Header and Footer
  - ☐ Pages and Paragraphs
- 



**Which of the following is an example of unstructured data?**

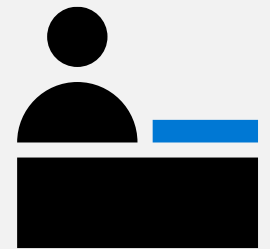
- ☐ An Employee table with columns Employee ID, Employee Name, and Employee Designation
  - ☒ Audio and Video files
  - ☐ A table within SQL Server database
- 



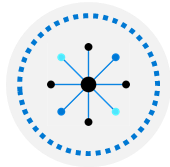
**What of the following is an example of a streaming dataset?**

- ☒ Data from sensor feeds
- ☐ Sales data for the past month
- ☐ List of employees working for a company

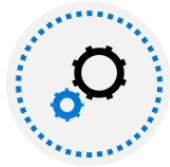
## Lesson 2: Explore roles and responsibilities in the world of data



## Lesson 2 objectives



Explore data job roles



Explore common tasks and tools for data job roles

# Roles in data

## Database Administrator

- Database Management
- Implements Data Security
- Backups
- User Access
- Monitors performance



## Data Engineer

- Data Pipelines and processes
- Data Ingestion storage
- Prepare data for Analytics
- Prepare data for analytical processing



## Data Analyst

- Provides insights into the data
- Visual Reporting
- Modeling Data for Analysis
- Combines data for visualization and analysis



# Common tools – Database administrator

## Azure Data Studio

Graphical interface for managing on-premises and cloud-based data services

Runs on Windows, macOS, Linux

## SQL Server Management Studio

Graphical interface for managing on-premises and cloud-based data services

Runs on Windows

Comprehensive Database Administration tool

## Azure Portal/CLI

Tools for management and provisioning of Azure Data Services

Manual and automation of scripts using Azure Resource Manager or Command Line Interface scripting

# Common tools – Data engineering

## Azure Synapse Studio

Azure Portal integrated to manage Azure Synapse

Data Ingestion (Azure Data Factory)

Management of Azure Synapse assets (SQL Pools/Spark Pool)

## SQL Server Management Studio

Graphical interface for managing on-premises and cloud-based data services

Runs on Windows

Comprehensive Database Administration tool

## Azure Portal/CLI

Tools for management and provisioning of Azure resources

Manual and automation of scripts using Azure Resource Manager or Command Line Interface scripting

# Common tools – Data analyst

## Power BI Desktop

Data Visualization tool  
Model and Visualize Data  
Management of Azure Synapse  
assets (SQL Pools/Spark Pool)

## Power BI Portal/ Power BI Service

Authoring and management of  
Power BI reports  
Authoring of Power BI dashboards  
Share Reports/Datasets

## Power BI Report Builder

Data Visualization tool for  
paginated reports  
Model and Visualize paginated  
reports



# Lesson 2: Knowledge check



**Which one of the following tasks is a role of a database administrator?**

- ☒ Backing up and restoring databases
  - ☐ Creating dashboards and reports
  - ☐ Identifying data quality issues
- 



**Which of the following tools is a visualization and reporting tool?**

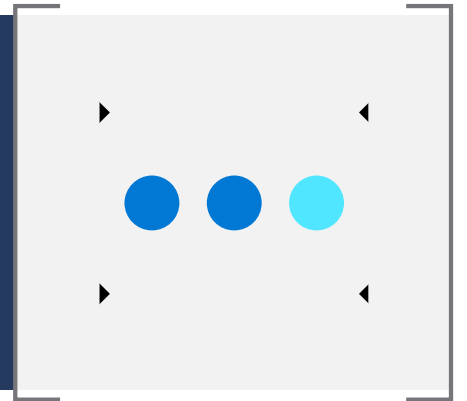
- ☐ SQL Server Management Studio
  - ☒ Power BI
  - ☐ SQL
- 



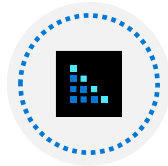
**Which one of the following roles is not a data job role?**

- ☒ Systems Administrator
- ☐ Data Analyst
- ☐ Database Administrator

## Lesson 3: Describe concepts of relational data



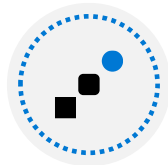
## Lesson 3 objectives



Explore the characteristics of relational data

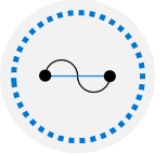


Define tables, indexes, and views



Explore relational data workload services in Azure

# Identify relational database use cases



## **IoT:**

Although typically considered for non-relational, the data from IoT devices could be structured and consistent

---



## **Online transaction processing:**

For example order systems that perform many small transactional updates

---



## **Data warehousing:**

Large amounts of data can be imported from multiple sources and structured to enable high-performance queries

# Tables

| Customers  |                 |               |
|------------|-----------------|---------------|
| CustomerID | CustomerName    | CustomerPhone |
| 100        | Muisto Linna    | XXX-XXX-XXXX  |
| 101        | Noam Maoz       | XXX-XXX-XXXX  |
| 102        | Vanja Matkovic  | XXX-XXX-XXXX  |
| 103        | Qamar Mounir    | XXX-XXX-XXXX  |
| 104        | Zhenis Omar     | XXX-XXX-XXXX  |
| 105        | Claude Paulet   | XXX-XXX-XXXX  |
| 106        | Alex Pettersen  | XXX-XXX-XXXX  |
| 107        | Francis Ribeiro | XXX-XXX-XXXX  |

Data is stored in a table

Table consists of rows and columns

All rows have same # of columns

Each column is defined by a datatype

# Entities

| Customers  |                |               |
|------------|----------------|---------------|
| CustomerID | CustomerName   | CustomerPhone |
| 100        | Muisto Linna   | XXX-XXX-XXXX  |
| 101        | Noam Maoz      | XXX-XXX-XXXX  |
| 102        | Vanja Matkovic | XXX-XXX-XXXX  |
| 103        | Qamar Mounir   | XXX-XXX-XXXX  |
| 104        | Zhenis Omar    | XXX-XXX-XXXX  |
| 105        | Claude Paulet  | XXX-XXX-XXXX  |
| 106        | Alex Pettersen | XXX-XXX-XXXX  |

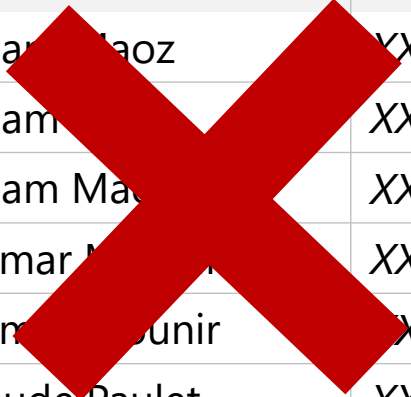
**An entity is a representation of an item which can be physical (such as a customer or a product), or virtual (such as an order).**

**Entities are connected by relations enabling interaction. For example, a customer can place an order for a product**

# Normalization

| Customers  |                |               |
|------------|----------------|---------------|
| CustomerID | CustomerName   | CustomerPhone |
| 100        | Muisto Linna   | XXX-XXX-XXXX  |
| 101        | Noam Maoz      | XXX-XXX-XXXX  |
| 102        | Vanja Matkovic | XXX-XXX-XXXX  |
| 103        | Qamar Mounir   | XXX-XXX-XXXX  |
| 104        | Zhenis Omar    | XXX-XXX-XXXX  |
| 105        | Claude Paulet  | XXX-XXX-XXXX  |
| 106        | Alex Pettersen | XXX-XXX-XXXX  |

| Orders  |               |               |
|---------|---------------|---------------|
| OrderID | CustomerName  | CustomerPhone |
| AD100   | Noam Maoz     | XXX-XXX-XXXX  |
| AD101   | Noam Maoz     | XXX-XXX-XXXX  |
| AD102   | Noam Maoz     | XXX-XXX-XXXX  |
| AX103   | Qamar Mounir  | XXX-XXX-XXXX  |
| AS104   | Qamar Mounir  | XXX-XXX-XXXX  |
| AR105   | Claude Paulet | XXX-XXX-XXXX  |
| MK106   | Muisto Linna  | XXX-XXX-XXXX  |



Data is normalized to:

Reduce storage

Avoid data duplication

Improve data quality

# Relations

| Customers  |                |               | Orders  |            |               |
|------------|----------------|---------------|---------|------------|---------------|
| CustomerID | CustomerName   | CustomerPhone | OrderID | CustomerID | SalesPersonID |
| 100        | Muisto Linna   | XXX-XXX-XXXX  | AD100   | 101        | 200           |
| 101        | Noam Maoz      | XXX-XXX-XXXX  | AD101   | 101        | 200           |
| 102        | Vanja Matkovic | XXX-XXX-XXXX  | AD102   | 101        | 200           |
| 103        | Qamar Mounir   | XXX-XXX-XXXX  | AX103   | 103        | 201           |
| 104        | Zhenis Omar    | XXX-XXX-XXXX  | AS104   | 103        | 201           |
| 105        | Claude Paulet  | XXX-XXX-XXXX  | AR105   | 105        | 200           |
| 106        | Alex Pettersen | XXX-XXX-XXXX  | MK106   | 105        | 201           |

## In a normalized database schema:

Primary Keys and Foreign keys are used to define relationships

No data duplication exists (other than key values in 3<sup>rd</sup> Normal Form (3NF))

Data is retrieved by joining tables together in a query



# Indexes

| Customers  |                |               |
|------------|----------------|---------------|
| CustomerID | CustomerName   | CustomerPhone |
| 100        | Muisto Linna   | XXX-XXX-XXXX  |
| 101        | Noam Maoz      | XXX-XXX-XXXX  |
| 102        | Vanja Matkovic | XXX-XXX-XXXX  |
| 103        | Qamar Mounir   | XXX-XXX-XXXX  |
| 104        | Zhenis Omar    | XXX-XXX-XXXX  |
| 105        | Claude Paulet  | XXX-XXX-XXXX  |
| 106        | Alex Pettersen | XXX-XXX-XXXX  |

| IDX-CustomerRegion |         |
|--------------------|---------|
| CustomerID         | Region  |
| 100                | France  |
| 101                | Brazil  |
| 102                | Croatia |
| 103                | Jordan  |
| 104                | Spain   |
| 105                | France  |
| 106                | USA     |

## An index:

Optimizes search queries for faster data retrieval

Reduces the amount of data pages that need to be read to retrieve the data in a SQL Statement

Data is retrieved by joining tables together in a query

# View

| Customers  |                |               |
|------------|----------------|---------------|
| CustomerID | CustomerName   | CustomerPhone |
| 100        | Muisto Linna   | XXX-XXX-XXXX  |
| 101        | Noam Maoz      | XXX-XXX-XXXX  |
| 102        | Vanja Matkovic | XXX-XXX-XXXX  |
| 103        | Qamar Mounir   | XXX-XXX-XXXX  |
| 104        | Zhenis Omar    | XXX-XXX-XXXX  |
| 105        | Claude Paulet  | XXX-XXX-XXXX  |
| 106        | Alex Pettersen | XXX-XXX-XXXX  |

| Orders  |            |               |
|---------|------------|---------------|
| OrderID | CustomerID | SalesPersonID |
| AD100   | 101        | 200           |
| AD101   | 101        | 200           |
| AD102   | 101        | 200           |
| AX103   | 103        | 201           |
| AS104   | 103        | 201           |
| AR105   | 105        | 200           |
| MK106   | 105        | 201           |
| DB205   | 100        | 205           |

Create the definition of a view:

```
CREATE VIEW
vw_customerorders AS
SELECT Customers.CustomerID,
Customers.CustomerName,
Orders.OrderID FROM
Customers JOIN Orders on
Customers.CustomerID =
Orders.CustomerID
```

Retrieve the orders placed by customer 102 using the view:

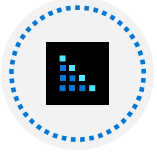
```
SELECT CustomerName, OrderID
from vw_customerorders WHERE
CustomerID=102
```

A view is a virtual table based on the result set of query:

Views are created to simplify the query

Combine relational data into a single pane view

# Lesson 3: Knowledge check



**Which one of the following statements is a characteristic of a relational database?**

- ☐ All data must be stored as character strings
  - ☒ A row in a table represents a single entity
  - ☐ Different rows in the same table can contain different columns
- 



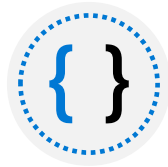
**What is an index?**

- ☒ A structure that enables you to locate rows in a table quickly, using an indexed value
- ☐ A virtual table based on the result set of a query
- ☐ A structure comprising rows and columns that you use for storing data

## Lesson 4: Explore concepts of non-relational data



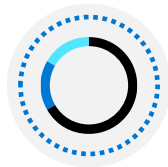
## Lesson 4 objectives



Explore the characteristics of non-relational data



Define types of non-relational data



Describe NoSQL, and the types of non-relational databases

# Explore characteristics of non-relational data

## Entities

```
## Customer 1 ID: 1
Name: Mark Hanson
Telephone: [ Home: 1-999-9999999, Business: 1-888-8888888, Cell: 1-777- 7777777 ]
Address: [ Home: 121 Main Street, Some City, NY, 10110,
           Business: 87 Big Building, Some City, NY, 10111 ]
## Customer 2 ID: 2
Title: Mr
Name: Jeff Hay
Telephone: [ Home: 0044-1999-333333, Mobile: 0044-17545-444444 ]
Address: [ UK: 86 High Street, Some Town, A County, GL8888, UK,
           US: 777 7th Street, Another City, CA, 90111 ]
```

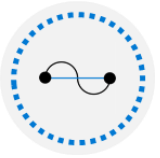
### Non-relational collections can have:

Multiple entities in the same collection or container with different fields

Have a different, non-tabular schema

Are often defined by labeling each field with the name it represents

# Identify non-relational database use cases



## **IoT and Telematics:**

Often require to ingest large amounts of data in frequent burst of activity, data is either semi structured or structured, often requires real time processing

---



## **Retail and Marketing:**

Common scenarios for globally distributed data, document storage

---



## **Gaming:**

In-game stats, social media integration, leaderboards, low-latency applications

---



## **Web and Mobile:**

Commonly used with web click analytics, modern applications including bots

# Types of non-relational data

## What is semi-structured data?

Data structure is defined within the actual data by fields. Format/file types include:



JSON

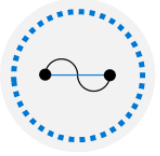
AVRO

ORC

Parquet



# What is unstructured data?



**Does not naturally contain fields:**

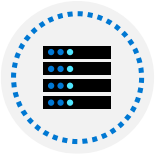
*Examples: video, audio, media streams, documents*

---



**Often used to extract data organization and categorize or identify “structures”**

---



**Frequently used in combination with Machine Learning or Cognitive Services capabilities to “extract data” by using:**

Text Analytics

Sentiment Analysis with Cognitive APIs

Vision API

# What is NoSQL?

Loose term, to describe non-relational



Key-value  
stores

Document  
based

Column  
family  
databases

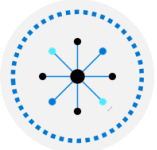
Graph  
Databases

# What is a graph database?

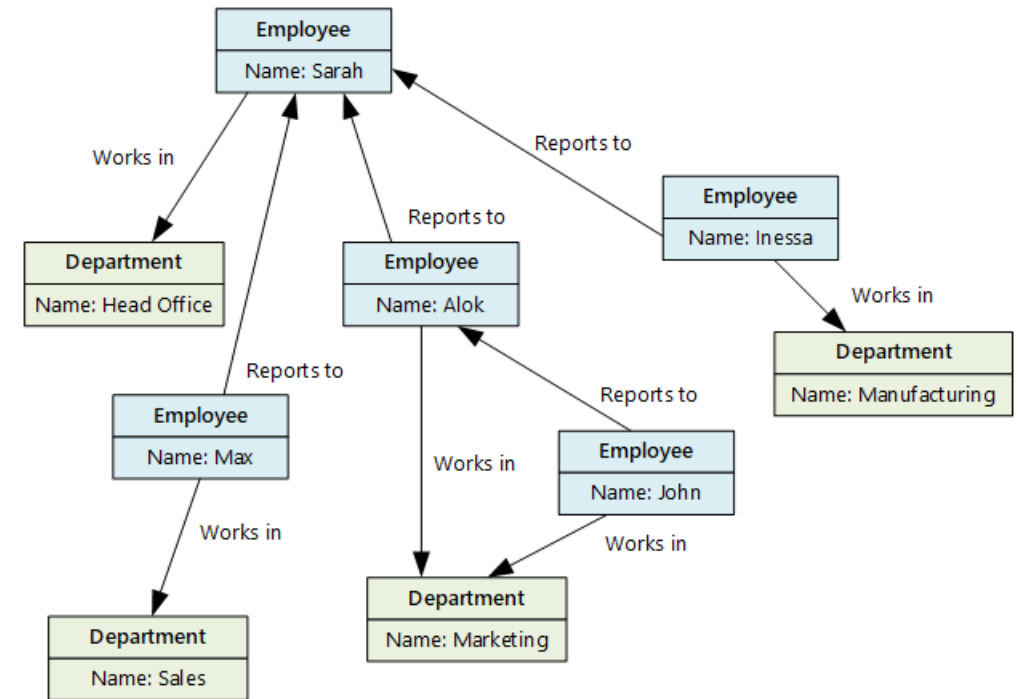


Stores entities centric around relationships

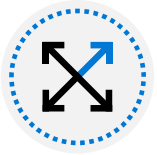
---



Enables applications to perform queries traversing a network of nodes and edges



# Lesson 4: Knowledge check



Which of the following services should you use to implement a non-relational database?

- ☒ Azure Cosmos DB
  - ☐ Azure SQL Database
  - ☐ The Gremlin API
- 



Which of the following is a characteristic of non-relational databases?

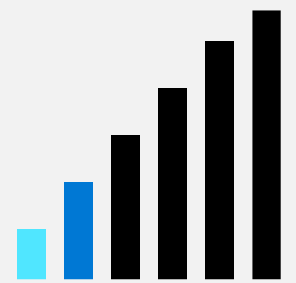
- ☐ Non-relational databases contain tables with flat fixed-column records
  - ☐ Non-relational databases require you to use data normalization techniques to reduce data duplication
  - ☒ Non-relational databases are either schema free or have relaxed schemas
- 



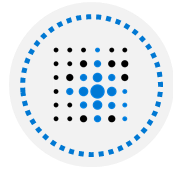
You are building a system that monitors the temperature throughout a set of office blocks, and sets the air conditioning in each room in each block to maintain a pleasant ambient temperature. Your system has to manage the air conditioning in several thousand buildings spread across the country or region, and each building typically contains at least 100 air-conditioned rooms. What type of NoSQL data store is most appropriate for capturing the temperature data to enable it to be processed quickly?

- ☒ A key-value store
- ☐ A column family database
- ☐ Write the temperatures to a blob in Azure Blob storage

## Lesson 5: Explore concepts of data analytics



# Lesson 5 objectives



Learn about data ingestion and processing



Explore data visualization

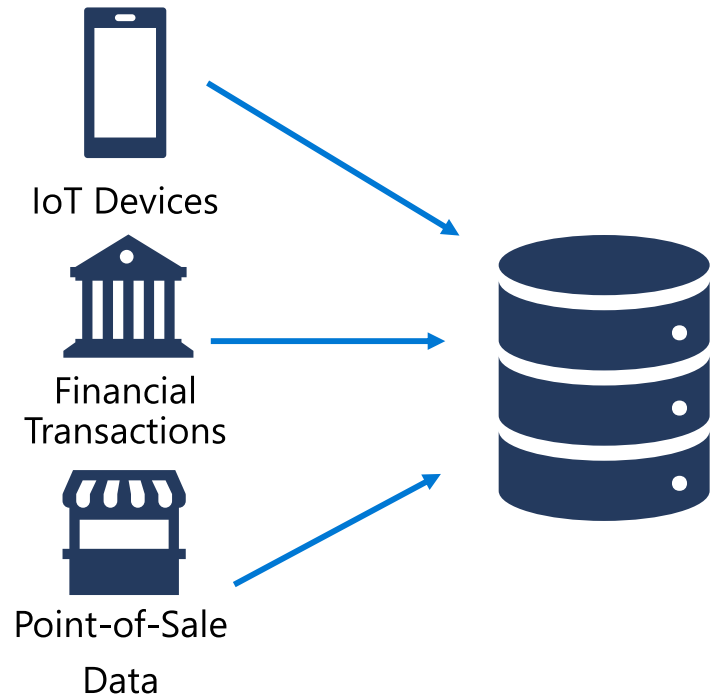


Explore data analytics

# The Data Journey

## Data Ingestion

The process of obtaining and importing data for immediate use or storage in a database



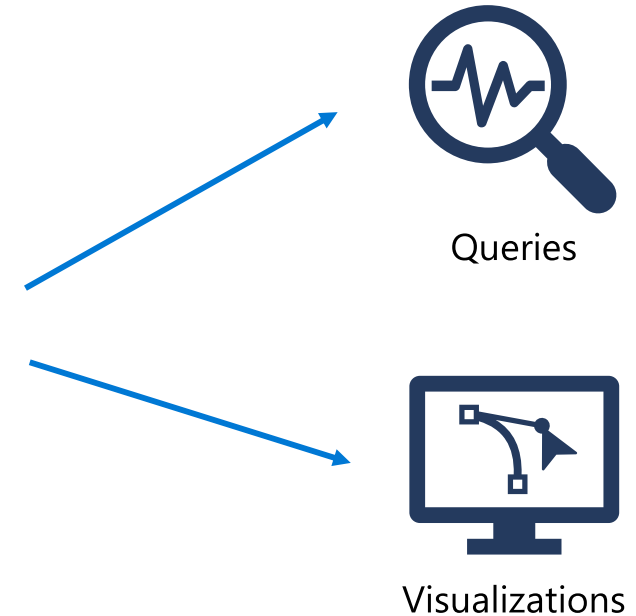
## Data Processing

Takes the data in its raw form, cleans it, and converts it into a more meaningful format



## Data Visualization

Query the data and create graphical representations of information and data



# Data visualization

A business model can contain an enormous amount of information – there are techniques to analyze and understand the information in your models



Reporting



Business intelligence (BI)



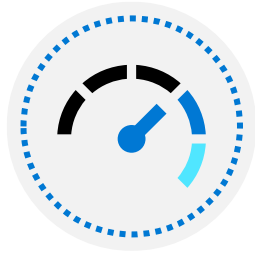
Data visualization



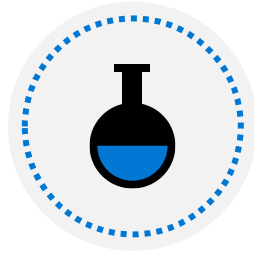
# Explore data analytics



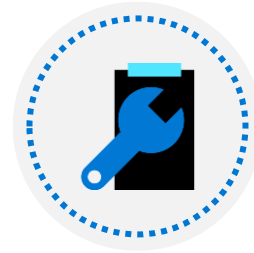
**Descriptive**



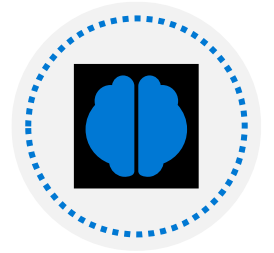
**Diagnostic**



**Predictive**



**Prescriptive**



**Cognitive**

# Lesson 5: Knowledge check



## What is data ingestion?

- ☐ The process of transforming raw data into models containing meaningful information
  - ☐ Analyzing data for anomalies
  - ☒ Capturing raw data streaming from various sources and storing it
- 



## Which one of the following visuals displays the major contributors to a selected result or value?

- ☒ Key influencers
  - ☐ Column and bar chart
  - ☐ Matrix chart
- 



## Which type of analytics helps answer questions about what has happened in the past?

- ☒ Descriptive analytics
- ☐ Prescriptive analytics
- ☐ Predictive analytics

