DP-900: Microsoft Azure Data Fundamentals

Mallik Gandhamsetty

Module 1: Core Data Concepts

1. Explore core data concepts
2. Explore roles and responsibilities in the world of data
3. Describe concepts of relational data
4. Explore concepts of non-relational data
5. Explore concepts of data analytics

Lesson 1: Core Data Concepts

- **What is Data?**
- **Data Classification - Structure**
  - Structured
  - Semi Structured
  - Unstructured
- **Data Classification - Usage**
  - Relational
  - Non-Relational
- **Azure Storage Services**
  - Azure SQL Database
  - Azure Cosmos DB
  - Azure Blob Storage
- **Data Processing Solutions**
  - OLTP
  - OLAP
- **Data Workloads**
  - Batch
  - Streaming

What is Data?

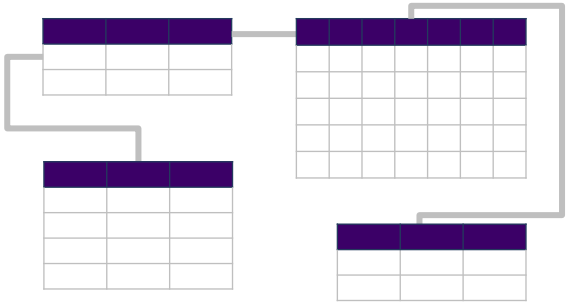
- Data is a *collection of facts* such as numbers, descriptions, and observations, objects etc

- Data can be collected, stored and processed in a variety of forms - *structured*, *semi-structured*, and *unstructured* forms

https://techjury.net/blog/how-much-data-is-created-every-day/

Structured Data

- Structured data is typically tabular data that is represented by rows and columns in a database.
- Example – Datawarehouse, ERP, CRM



Structured

Conforms to a schema

Order	CustID	Month	Item	Color	Price
101	20051	Dec	Pen	Red	2.99
102	20045	Mar	Pencil	Blue|Yellow|Red	3.99
103	29584	May	Eraser	Blue	1.25
104	29584	May	Pen	White	2.25
105	29584	May	Pencil	Blue|Yellow|Red	2.99
106	27485	Jan	Eraser	Blue|Yellow	2.75
107	29574	Jan	Marker	Green	1.75
108	24447	Feb	Marker	Yellow|Blue	7.25
109	26466	Jul	Pen	Black|Red	5.25
110	27467	Jun	Pencil	Black	2.95

Semi-Structured Data

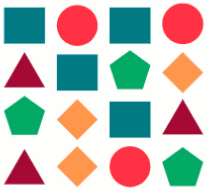
- Data that has some sort of structure but doesn't fit in a relational database.
- Example: Documents held in *JavaScript Object Notation* (JSON) or XML formats

```
## 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"
}
```

Semi-structured

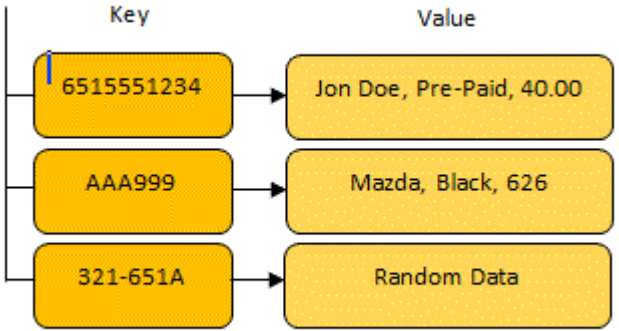
Some level of organization



```
▼<div class="new-main-menu">
  ▼<div class="header-desktop-block">
    ▼<div class="container new-menu">
      ▶<a class="main-logo" rel="home" href="https://
      databricks.com/ " title="Databricks">…</a>
      ▼<div id="new-m" class="menu-bar">
        ▼<div id="mega-menu-wrap-headerNew" class="mega-menu-
        wrap">
          ▶<div class="mega-menu-toggle">…</div>
          ▼<ul id="mega-menu-headerNew" class="mega-menu max-
          mega-menu mega-menu-horizontal" data-event=
          "hover_intent" data-effect="fade_up" data-effect-speed=
          "200" data-effect-mobile="disabled" data-effect-speed-
          mobile="0" data-panel-width="body" data-panel-inner-
          width="#new-m" data-mobile-force-width="false" data-
          second-click="close" data-document-click="collapse"
          data-vertical-behaviour="standard" data-breakpoint=
          "1199" data-unbind="true">
            ▶<li class="mega-main-bar-li mega-menu-item mega-
            menu-item-type-custom mega-menu-item-object-custom
            mega-menu-item-has-children mega-menu-megamenu mega-
```

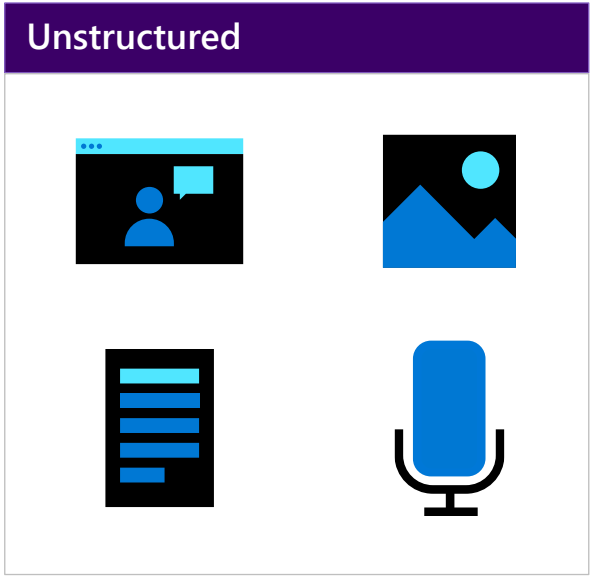
Semi-Structured Data

Semi-structured data can be stored in *key-value* stores or *graph* databases as well



Unstructured Data

- Unstructured data is data which is not organized in any predefined manner.
- Example – audio and video files, and binary data files



Data Storage Services in Azure

- Structured Data - **Azure SQL Database**
- Semi-Structured Data - **Azure Cosmos DB**
- Unstructured Data - **Azure Blob Storage**

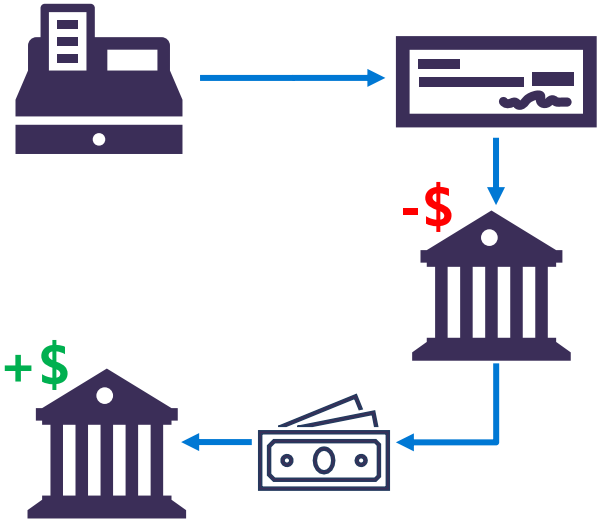
Data Centric Solutions

- Transactional System (OLTP)
- Analytical System (OLAP)

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 Solutions

Online Transactional Processing (OLTP)

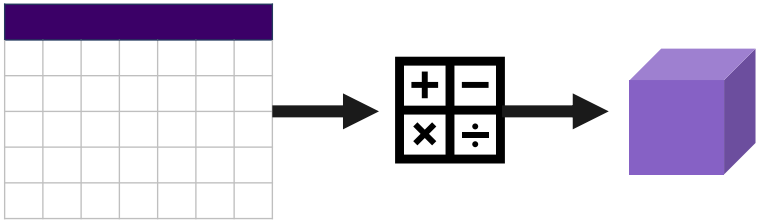
Customer		
CustomerID	CustomerName	CustomerPhone

Orders		
OrderID	CustomerID	OrderDate

Data is stored one transaction at a time. Day-to-day handling of transactions that result from enterprise operations

- Small, discrete, unit of work
- Often high-volume
- Data processed very quickly.

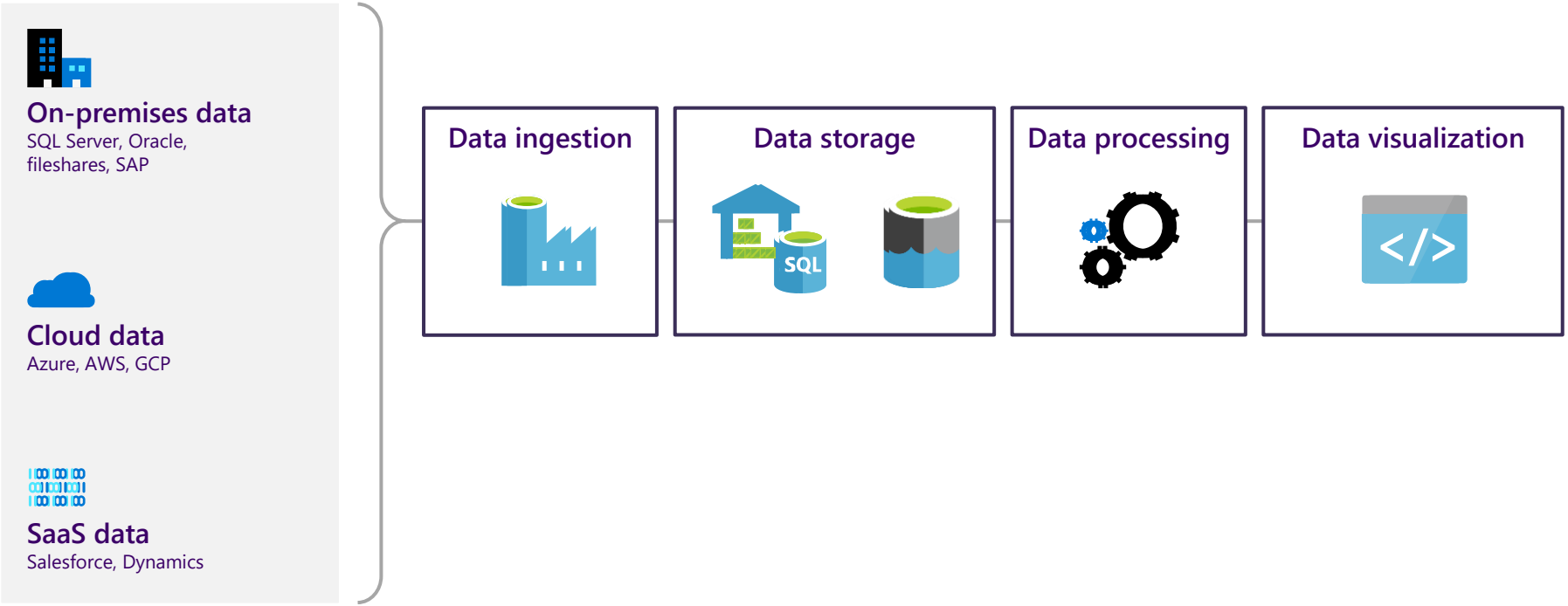
Online Analytical Processing (OLAP)



Data is periodically loaded, aggregated and stored in a cube. Analysis of information in a database for the purpose of making management decisions

- Big picture view of the information held in a database.
- Generate insights to make business decisions

Analytical System



Data Processing

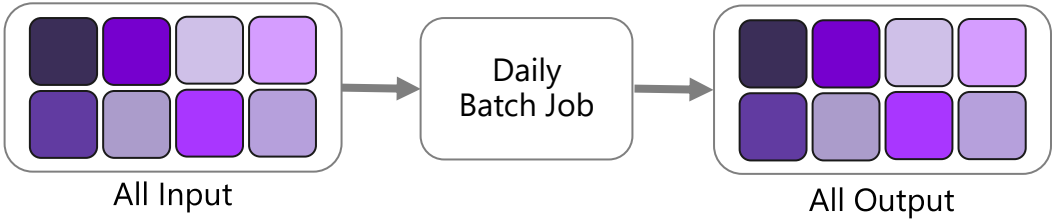
- Batch
- Streaming

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

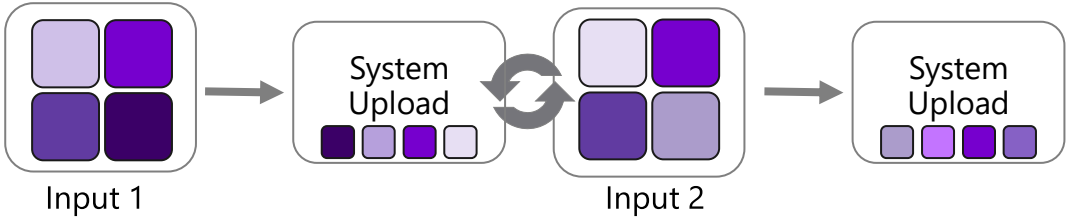
Example – Credit Card Bill



Stream Processing: handles data in real time. Each new piece of data is processed as and when it arrives.

ideal for time-critical operations that require an instant real-time response.

Example – Stock market, Heat alarm system, YouTube, Netflix



Batch vs Stream Processing

	Advantages	Disadvantages
Batch Processing	Large volumes of data can be processed at a convenient time.	High latency between ingesting the data and getting the results.
	Better resource utilization by running at idle time	Minor data errors can affect the whole batch
Stream Processing	No or Low latency between event occurrence and result computation	Can only process small volume of data, in real time

Batch vs Stream Processing

	Batch	Streaming
Data Scope	Can process all the data in the dataset in one go	Can only access the most recent data received
Data Size	Suitable for handling large datasets efficiently	intended for individual records or micro batches
Performance	latency of few hours	latency of few seconds/milliseconds
Analysis	used to perform complex analytics.	used for simple response functions, aggregates, or calculations such as rolling averages.

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

Module 1: Core Data Concepts

1. Explore core data concepts
2. Explore roles and responsibilities in the world of data
3. Describe concepts of relational data
4. Explore concepts of non-relational data
5. Explore concepts of data analytics

Data Job Roles

- Data job roles
- Common tasks and tools

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
- Notebook Support

SQL Server Management Studio

- Graphical interface for managing on-premises and cloud-based data services
- Runs on Windows
- Comprehensive Database Administration tool
- No Notebook like capability

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

Module 1: Core Data Concepts

1. Explore core data concepts
2. Explore roles and responsibilities in the world of data
3. Describe concepts of relational data
4. Explore concepts of non-relational data
5. Explore concepts of data analytics

Relational Data Concepts

- RDBMS Usage
- Characteristics
- Normalization
- Tables, Views, Indexes

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

Characteristics of Relational Data

Customers

Customer ID	Customer Name	Customer Address
C1	Fred	...
C2	Bert	...
C3	Jane	...

Products

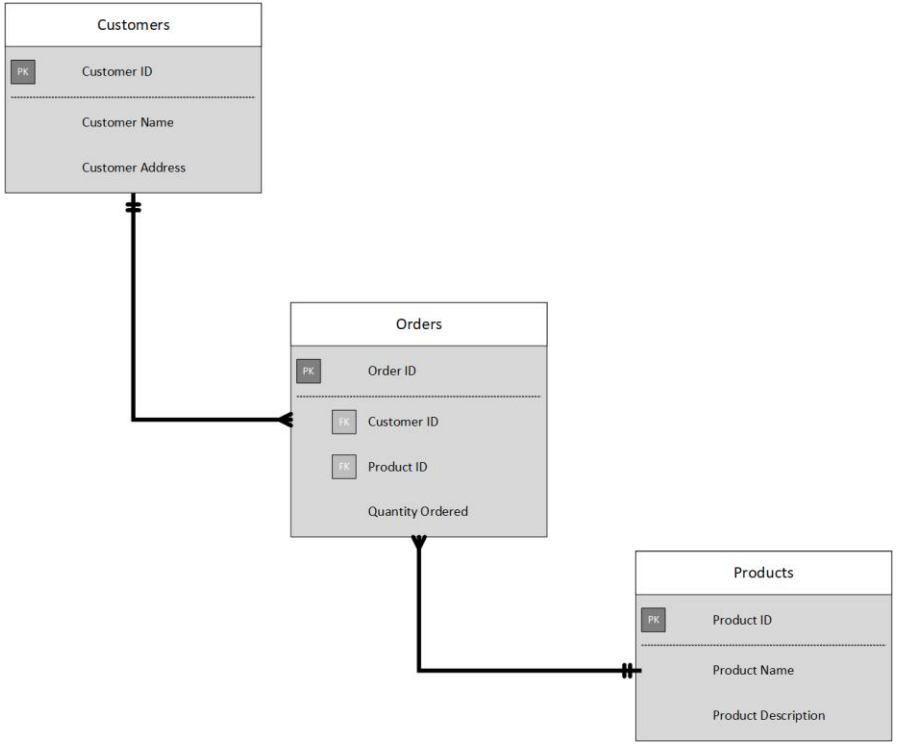
Product ID	Product Name	Description
P1	Shirt	...
P2	Tie	...
P3	Collar	...

Orders

Order ID	Customer ID	Product ID	Quantity
1000	C1	P1	1
1001	C2	P1	3
1002	C1	P3	1
1003	C1	P3	2
1004	C2	P2	4
1005	C1	P2	2
1006	C3	P3	1

- Data is stored in structures called ***Tables***
- Table consist of ***rows*** and ***columns***
- Each row represents a single instance of an entity
- Columns define the properties of the Entity
- Each column is defined by a ***datatype***
- All rows have the same number of columns

Characteristics of Relational Data



- Some columns are used to maintain ***relationships*** between tables
- Model shows the structure of the entities
- A ***Primary Key*** uniquely identifies each row
- A ***Foreign Key*** reference is a link to the primary key of another table
- These are used to maintain relationships between tables.

ANOMALIES

- Bad Table Designs cause redundant data to be stored and lead to problems known as anomalies.
- Redundancy means duplication of the data.

Course _no	Tutor	Room	Room_size	En_limit
353	Smith	A532	45	40
351	Smith	C320	100	60
355	Clark	H940	400	300
456	Turner	H940	400	45

Insert Anomaly

- An Insert Anomaly occurs **when certain attributes cannot be inserted** into the database without the presence of other attribute
- Say we built a new room (e.g. B123) and it has not yet been timetabled for any courses or members of staff. Can we insert it into this

Delete Anomaly

- A Delete Anomaly exists **when certain attributes are lost because of the deletion of other attributes**
- Say we wish to delete course_no 351 from the above table, but if we do so, the details of room also C320 get deleted as a side effect

Update Anomaly

- An Update Anomaly exists **when one or more instances of duplicated data is updated**, but not all.
- Say, Room H940 has been improved, it is now of Room_Size = 500. If we wish to update the size, we end up updating all other rows where room=H940.

Normalization

Data is normalized to:

Reduce storage	Avoid data duplication	Improve data quality

In a normalized database schema:

Primary Keys and Foreign keys are used to define relationships	No data duplication exists (other than key values in 3rd Normal Form (3NF)	Data is retrieved by joining tables together in a query

Normalization

Step 0: Unnormalized table

Student#	Trainer	TrainerRoom	Class1	Class2	Class3
1022	Jones	412	101-07	143-01	159-02
4123	Smith	216	101-07	143-01	179-04

1st NF Rules:

- Each table has a **primary key**: minimal set of attributes which can uniquely identify a record
- The values in each column of a table are atomic (**No multi-value** attributes allowed).
- There are **no repeating groups**: two columns do not store similar information in the same table

Normalization

Step 1: First normal form: No repeating groups

Student# %	Trainer	TrainerRoom	Class#
1022	Jones	900	101-07
1022	Jones	900	143-01
1022	Jones	900	159-02
4123	Smith	203	101-07
4123	Smith	203	143-01
4123	Smith	203	179-04

2nd NF Rules:

- All requirements for 1st NF must be met.
- Redundant data across multiple rows of a table must be moved to a separate table and should relate using ***Foreign Keys***

Normalization

Step 2: Second normal form: Eliminate redundant data

Student

Student#	Trainer	TrainerRoom
1022	Jones	900
4123	Smith	203

Registration

Student#	Class#
1022	101-07
1022	143-01
1022	159-02
4123	101-07
4123	143-01
4123	179-04

Normalization

Step 3: Third normal form: Eliminate data not dependent on key

Student

Student#	Trainer
1022	Jones
4123	Smith

Trainer

Trainer	Room#
Jones	900
Smith	203

Registration

Student#	Class#
1022	101-07
1022	143-01
1022	159-02
4123	101-07
4123	143-01
4123	179-04

Normalization - Example 2

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	Noa aoz	XXX-XXX-XXXX
AD101	Noam	*XXX-XXX-XXXX*
AD102	Noam Ma	*XXX-XXX-XXXX*
AX103	Qamar	*XXX-XXX-XXXX*
AS104	Qam unir	XX-XXX-XXXX
AR105	Claude Paulet	*XXX-XXX-XXXX*
MK106	Muisto Linna	*XXX-XXX-XXXX*

Table Relationships

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	101	201
AS104	103	201
AR105	105	200
MK106	105	201

Indexes

```
SELECT OrderID, ProductID
FROM Orders
WHERE CustomerID = "C1"
```

Index

Customer ID
C1
C1
C1
C1
C2
C2
C3

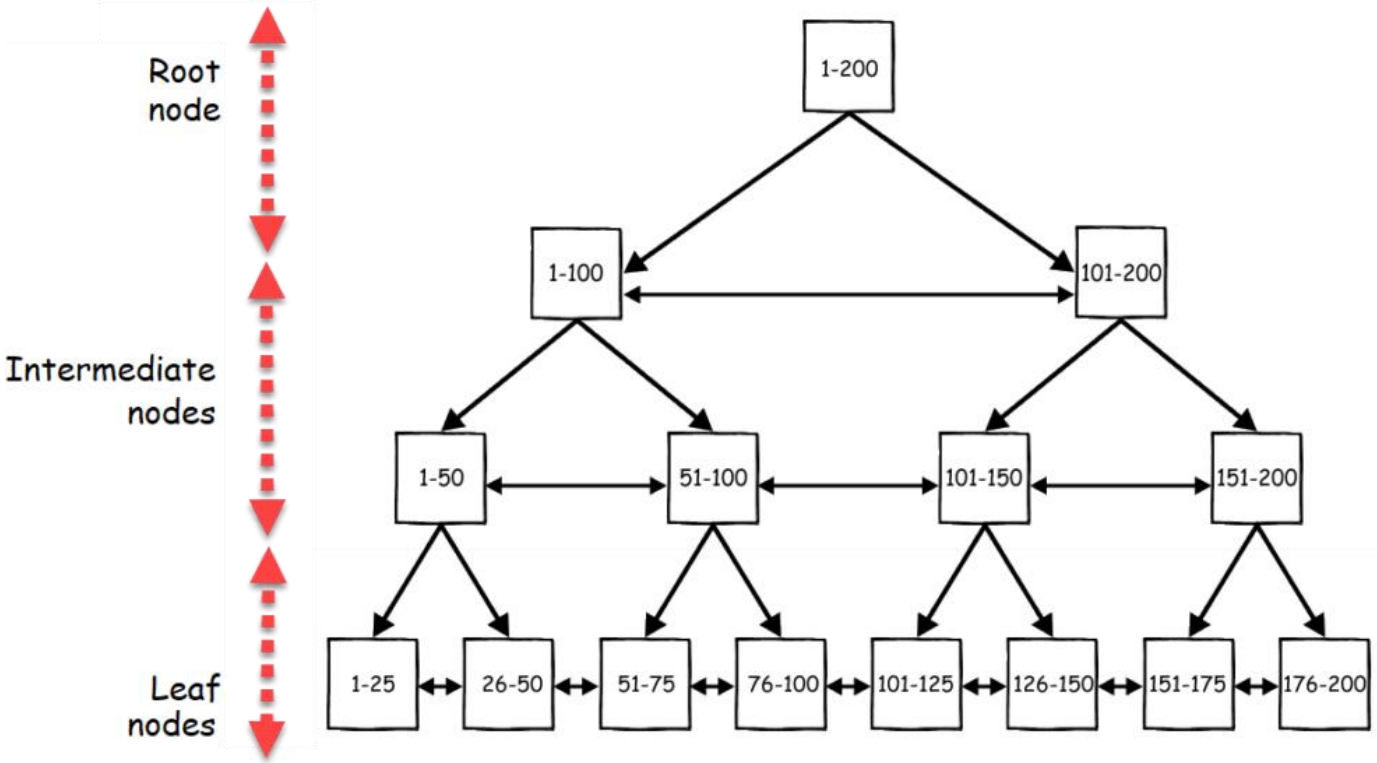
Orders

Order ID	Customer ID	Product ID	Quantity
1000	C1	P1	1
1001	C2	P1	3
1002	C1	P3	1
1003	C1	P3	2
1004	C2	P2	4
1005	C1	P2	2
1006	C3	P3	1

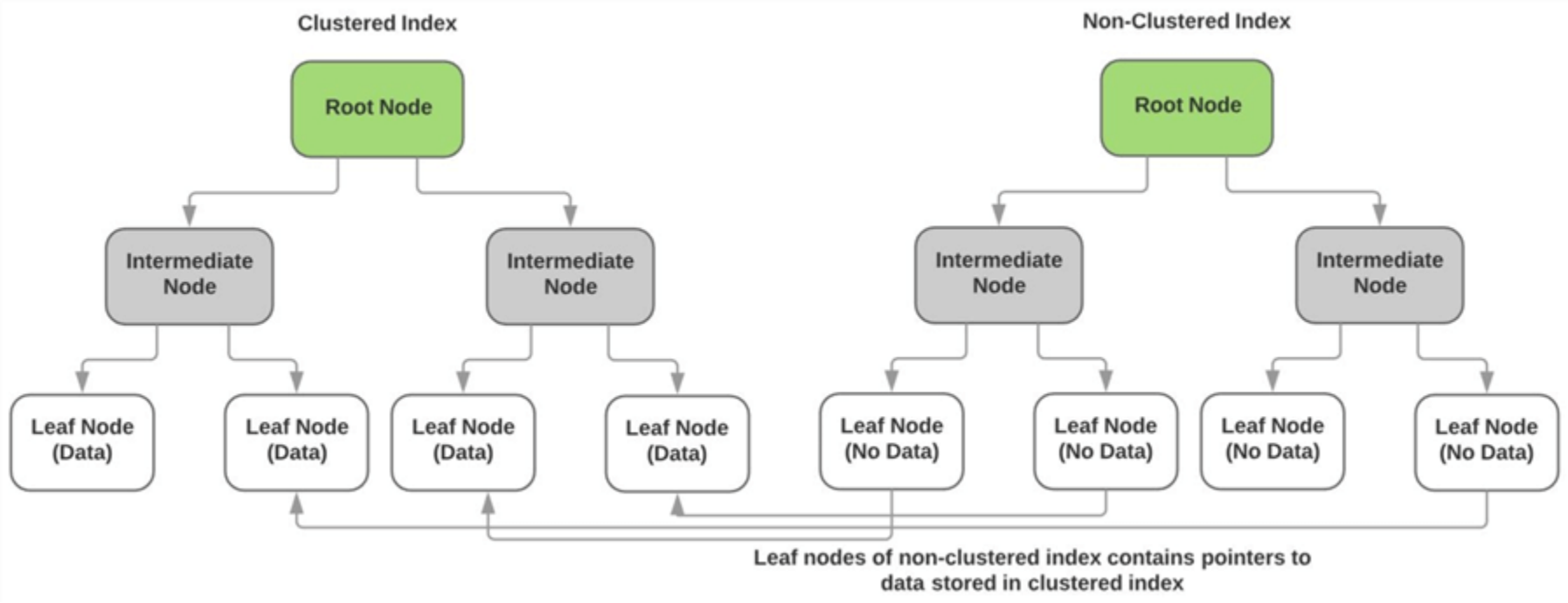
An index:

- 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

Indexes


Root node
Intermediate nodes
Leaf nodes
1-200
1-100
101-200
1-50
51-100
101-150
151-200
1-25
26-50
51-75
76-100
101-125
126-150
151-175
176-200

Indexes


Clustered Index
Root Node
Intermediate Node
Intermediate Node
Leaf Node (Data)
Leaf Node (Data)
Leaf Node (Data)
Leaf Node (Data)
Non-Clustered Index
Root Node
Intermediate Node
Intermediate Node
Leaf Node (No Data)
Leaf Node (No Data)
Leaf Node (No Data)
Leaf Node (No Data)
Leaf nodes of non-clustered index contains pointers to data stored in clustered index

Views

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

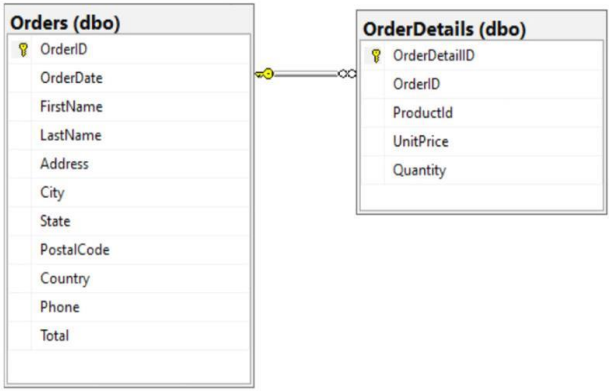
Module 1: Core Data Concepts

1. Explore core data concepts
2. Explore roles and responsibilities in the world of data
3. Describe concepts of relational data
4. Explore concepts of non-relational data
5. Explore concepts of data analytics

Non-Relational Data Concepts

- Characteristics
- Usage
- Types of Non-Relational Data
- Types of Non-Relational Databases

Explore characteristics of non-relational data




```
{
    "OrderId": 1,
    "OrderDate": 1574161910220,
    "FirstName": "John",
    "LastName": "Smith",
    "Address": "10 Street",
    "City": "City",
    "State": "VA",
    "OrderDetails": [
        {
            "UnitPrice": 7.99,
            "OrderDetailId": 2,
            "Quantity": 1,
            "ProductId": 259694,
            "OrderId": 1
        },
        {
            "UnitPrice": 7.99,
            "OrderDetailId": 3,
            "Quantity": 1,
            "ProductId": 295693,
            "OrderId": 1
        }
    ],
    "id": "795c50dc-1a83-11ea-bf07-00163ee85f66",
    "_rid": "VdgtAK23OMANAAAAAAAAAA==",
    "_self": "dbs/VdgtAA==/colls/VdgtAK23OMA=/docs/VdgtAK23OMANAAAAAAAAAA==/"
    "_etag": "\"370017e1-0000-1100-0000-5df770f20000\"",
    "_attachments": "attachments/",
    "_ts": 1576497394
}
```


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

What is NoSQL?

Loose term, to describe non-relational

Key-value stores

Key	Value
Bob	(123) 456-7890
Jane	(234) 567-8901
Tara	(345) 678-9012
Tiara	(456) 789-0123

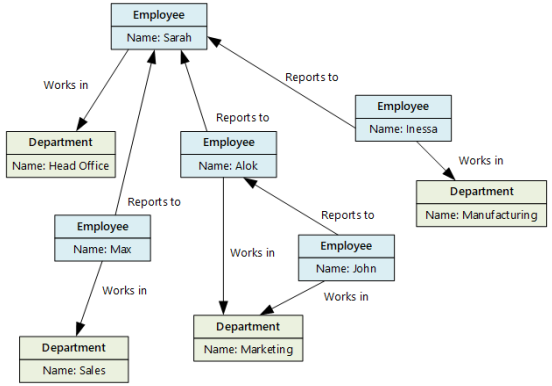
Column family databases

KEY	COLUMN FAMILIES	
ID	CUSTOMERINFO	ADDRESSINFO
1001	FirstName: Tom MiddleName: T LastName: Tester	Address1: 2001 Bayfront Dr. Address2: Suite#813 City: Tampa State: FL Zip: 34637 Country: US
1002	FirstName: Bob MiddleName: B LastName: Builder	Address1: 1234 Sunny Circle City: Beverly Hills State: CA Zip: 90210

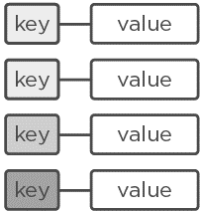
Document based

```
{
  "orderid": 12212,
  "orderdate": "12/4/2020",
  "customer":
      { "name": "Bob Smith", "email": "bobsmith@email.bob" },
  "status": "in process",
  "paymentmethod": "invoice",
  "products": [
      { "name": "Product 1", "quantity": 1 },
      { "name": "Product 2", "quantity": 1, status: 3 }
  ]
}
```

Graph Databases



Key-value stores



Key	Value
Bob	(123) 456-7890
Jane	(234) 567-8901
Tara	(345) 678-9012
Tiara	(456) 789-0123

What is a Key Value Store?

- Uses a simple key/value to store data
- Quick to query due to its simplicity
- Value can be JSON, BLOB, String etc.

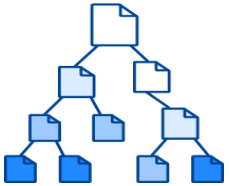
Use Cases:

- User profiles and session info on a website, blog comments, telecom directories, IP forwarding tables, shopping cart contents on e-commerce sites, and more.

Examples

- Cosmos DB Table API, Redis, Table Storage, Oracle NoSQL Database, Voldemorte, Aerospike, Oracle Berkeley DB

Document stores



```
{
 "orderid": 12212,
 "orderdate": "12/4/2020",
 "customer":
     { "name": "Bob Smith", "email": "bobsmith@email.bob" },
 "status": "in process",
 "paymentmethod": "invoice",
 "products": [
     { "name": "Product 1", "quantity": 1 },
     { "name": "Product 2", "quantity": 1, status: 3 }
 ]
}
```

What is a Document Datastore?

- Document-oriented model to store data
- Similar to key/value store, difference is that, the value in a document store database consists of semi-structured data.
- Each record and its associated data within a single document.
- Document stores are usually XML, JSON, BSON, YAML, etc.

Use Cases:

- Content management systems, blogging platforms, and other web applications, blog comments, chat sessions, tweets, ratings, etc.

Examples

- Cosmos DB, MongoDB, DocumentDB, CouchDB, MarkLogic, OrientDB

Column Family Databases

CustomerID	Column Family: Identity
001	**First name:** Mu Bae **Last name:** Min
002	**First name:** Francisco **Last name:** Vila Nova **Suffix:** Jr.
003	**First name:** Lena **Last name:** Adamcyz **Title:** Dr.

CustomerID	Column Family: Contact Info
001	**Phone number:** 555-0100 **Email:** someone@example.com
002	**Email:** vilanova@contoso.com
003	**Phone number:** 555-0120

What is a Column Family Datastore?

- Stores data using a column -oriented model
- Columns in each row are contained within that row
- Each row can have different columns to the other rows.
- Extremely quick to load and query

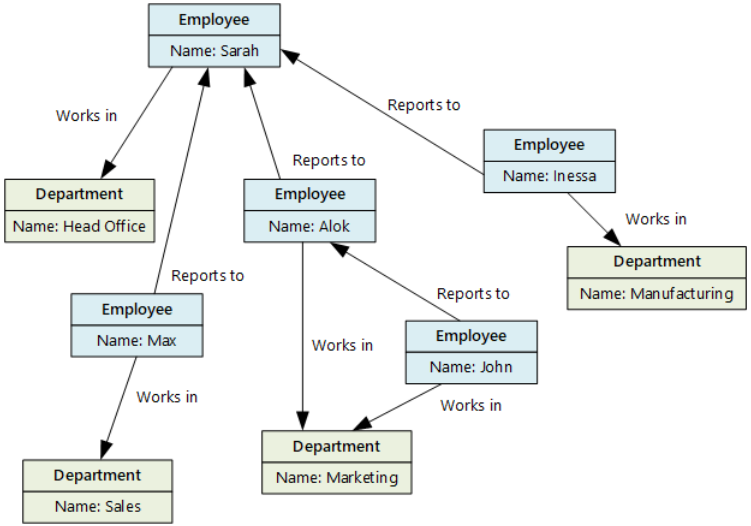
Use Cases:

- Sensor Logs [Internet of Things (IOT)], User preferences, Geographic information, Reporting systems, Time Series Data, Logging and other write heavy applications

Examples

- Cosmos DB, Bigtable, Cassandra, Hbase, Vertica, Druid, Accumulo, Hypertable

Graph Database



What is a Graph Datastore?

- Stores entities centric around relationships
- Enables applications to perform queries traversing a network of nodes and edges
- A **node** is a specific entity or piece of information
- **Edge** simply specifies the relationship between two nodes.
- Enables applications to perform queries traversing a network of nodes and edges

Use Cases:

- Social networks, real-time product recommendations, network diagrams, fraud detection, access management, and more

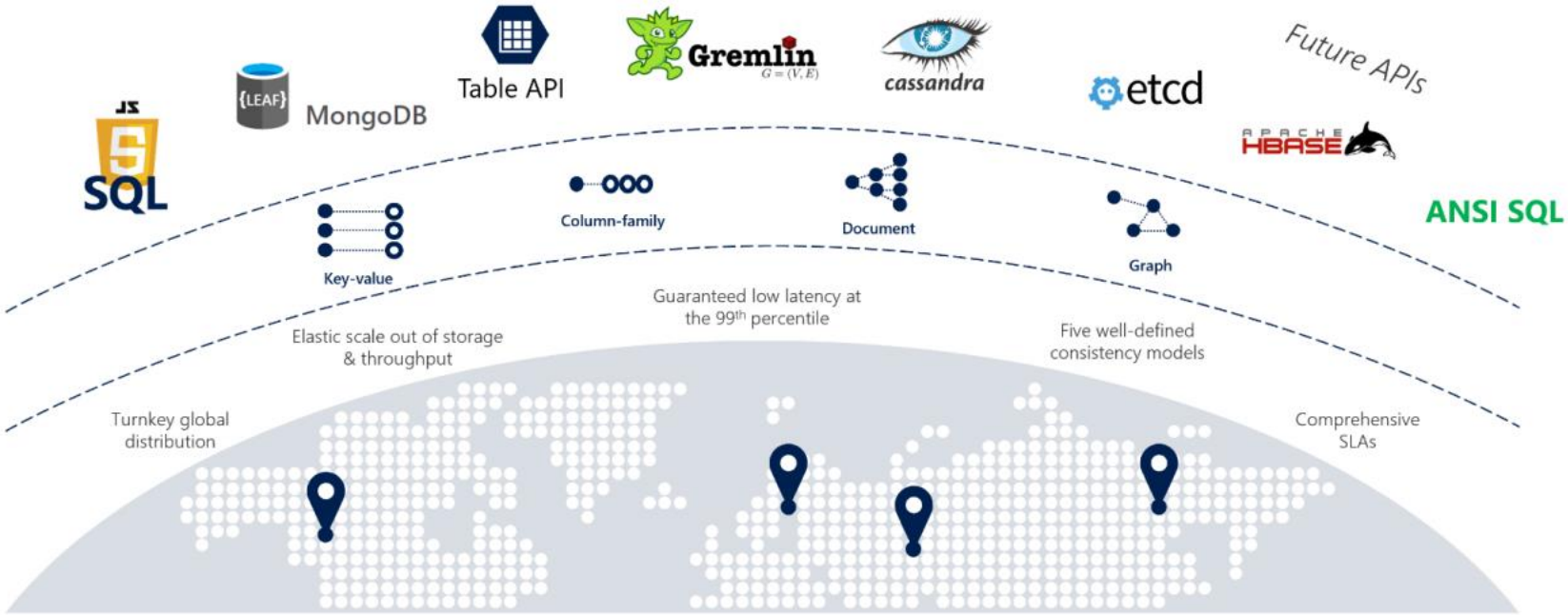
Examples

- Cosmos DB Gremlin API, Neo4j, Blazegraph, and OrientDB.

Azure No SQL Offerings

Azure Cosmos DB

Microsoft's globally distributed, massively scalable, multi-model database service



Lesson 4: Knowledge check

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

- ✔ 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

Module 1: Core Data Concepts

1. Explore core data concepts
2. Explore roles and responsibilities in the world of data
3. Describe concepts of relational data
4. Explore concepts of non-relational data
5. Explore concepts of data analytics

Data Analytics Concepts

- Data Ingestion
- Data Processing
- Data Visualization
- Data Analytics

The Data Journey

Data Ingestion

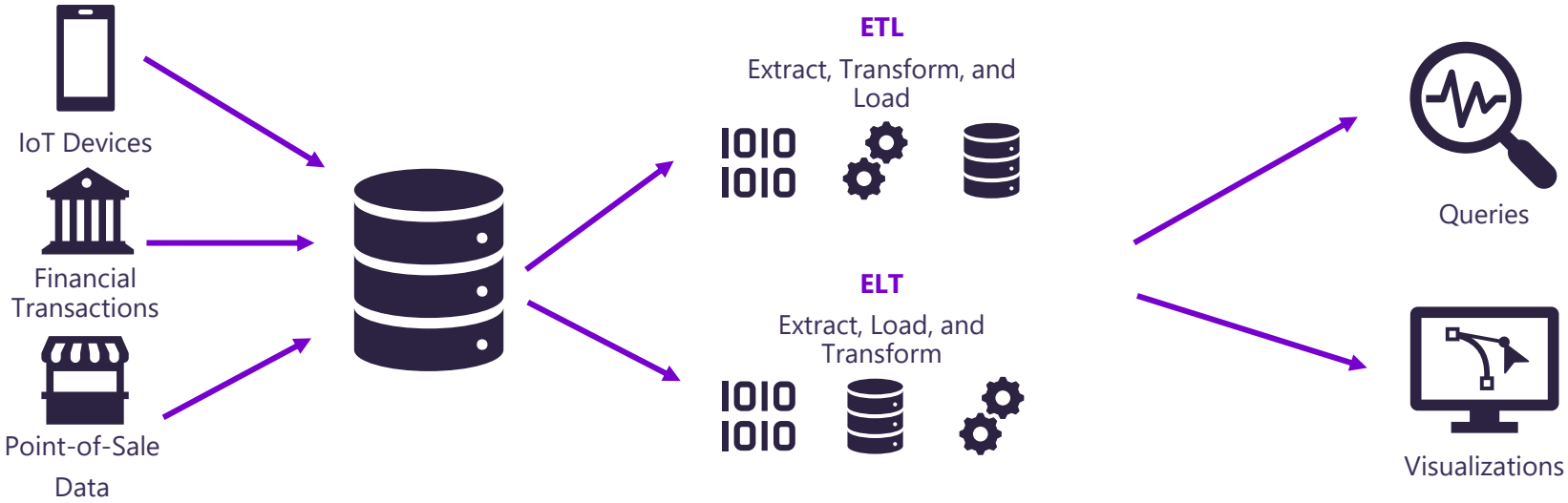
The process of obtaining and importing data for immediate use or storage in a database

Data Processing

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

Data Visualization

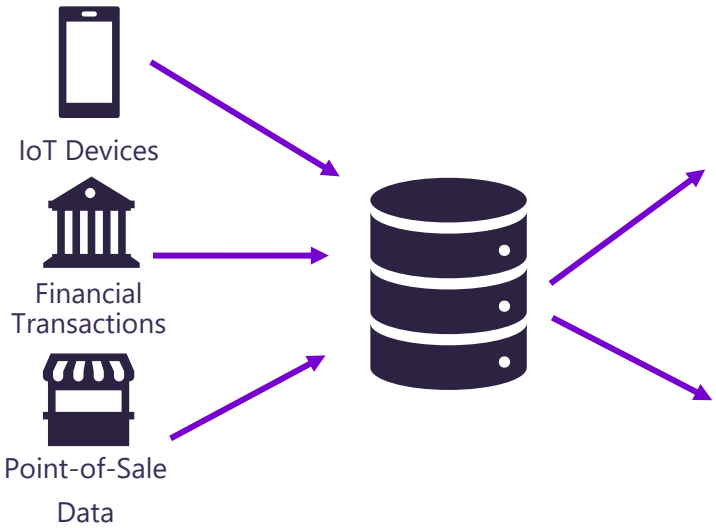
Query the data and create graphical representations of information and data



Data Ingestion

Data Ingestion

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



- Data ingestion is the process of ***extracting*** and ***importing*** data
- Data can arrive as a continuous ***stream*** or ***batches***
- Raw data can be stored at DBMS, a set of files, or some other type of fast, easily accessible storage.
- Ingestion process might perform:

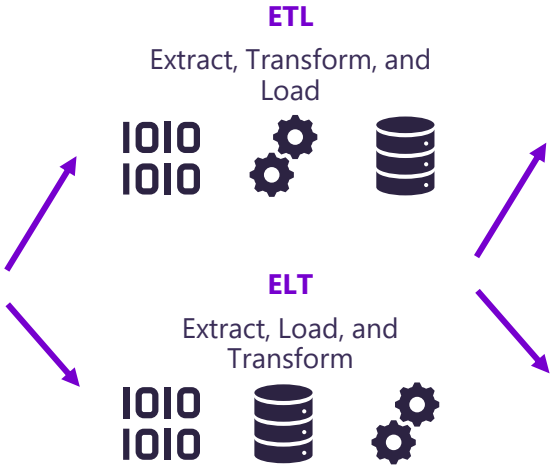
  Filtering: Example reject suspicious, corrupt, or duplicated data

  Simple transformations: converting data into a standard form. Example: reformat all date and time

Data Processing

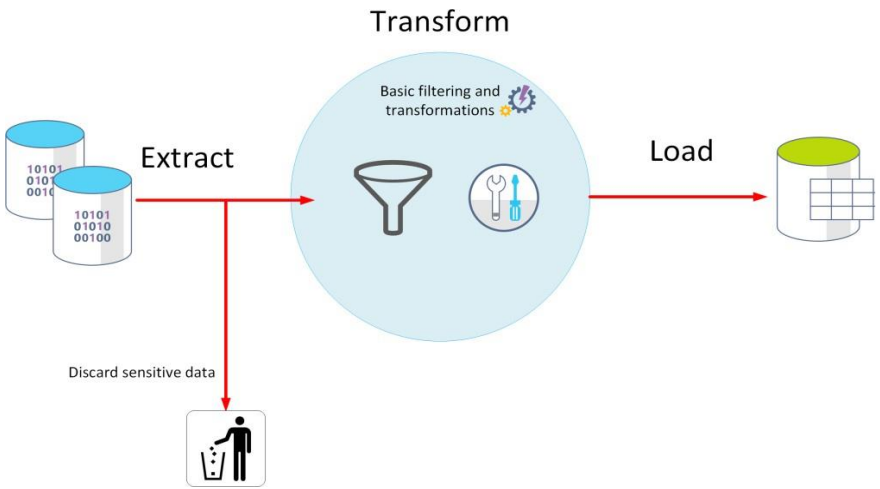
Data Processing

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



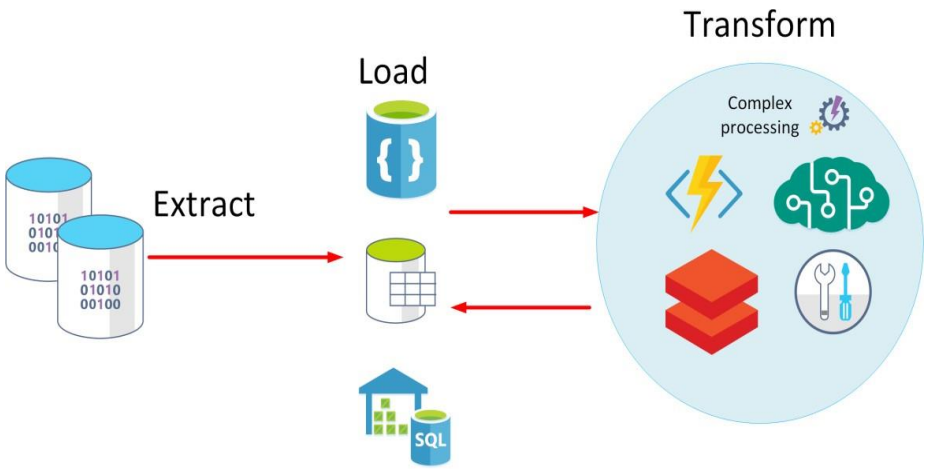
- Data processing takes the data in its raw form, cleans it, and converts it into a more meaningful format (tables, graphs, documents, and so on)
- The output of data processing is used to perform queries and generate visualizations
- **Data Cleaning:** removing anomalies, and applying filters and transformations
- **Data Wrangling:** capturing, filtering, cleaning, combining, and aggregating data

ETL - Extract, Transform, and Load



- Raw data is retrieved and transformed before being saved
- Suitable for systems that only require simple models
- Basic data cleaning tasks, deduplicating data, and reformatting the contents of individual fields.
- Stream-oriented approach - emphasis on throughput
- ETL can help with data privacy and compliance, removing sensitive data before it arrives in your analytical data models.
- Performed as a continuous pipeline of operations
- Example: SSIS, Informatica, DataStage, Ab Initio etc

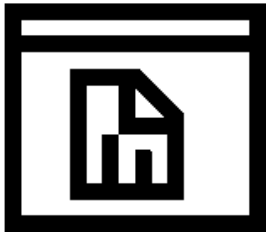
ELT - Extract, Load and Transform



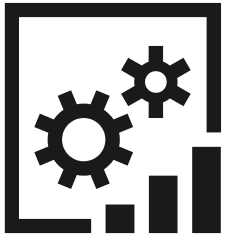
- Data is stored before being transformed
- More suitable for constructing complex models
- Iterative approach, often using periodic batch processing.
- Suitable where Target datastore is powerful enough to perform complex transformations
- Example: Azure Data Factory ingests data into Data Lake or Synapse Analytics and Compute services such as Azure HDInsight Hadoop, Azure Databricks or Synapse SQL/Spark Pools transform the data

Data visualization

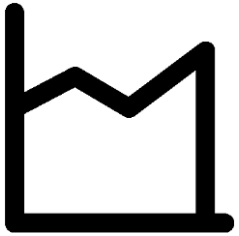
A business model can contain an enormous amount of information – there are techniques to analyze and understand the information in your models



Reporting



Business intelligence (BI)

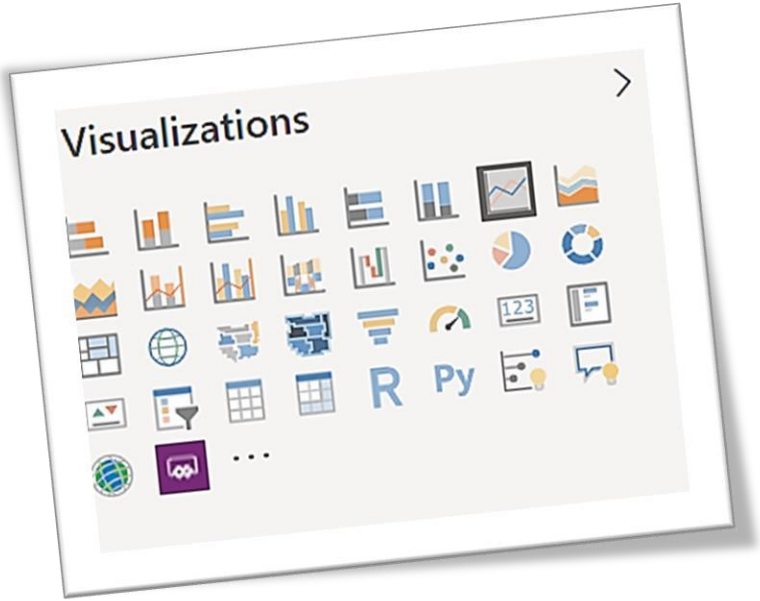


Data visualization

Data Visualization

What is Data Visualization?

- ***Graphical*** representation of information and data
- Using visual elements like ***charts***, ***graphs***, and ***maps***
- Helps you to focus on the meaning of data, rather than looking at the data itself
- Data visualization tools provide an accessible way to spot and understand ***trends***, ***outliers***, and ***patterns*** in data.
- In Azure we use **Power BI**



Data Visualization

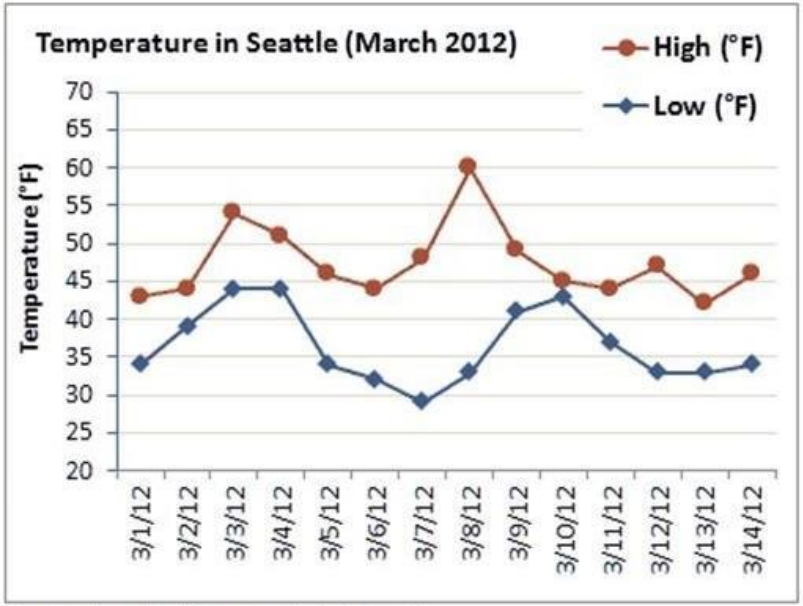
	A	B	C
1	Temperature Data for Seattle		
2	Date	High (°F)	Low (°F)
3	3/1/12	43	34
4	3/2/12	44	39
5	3/3/12	54	44
6	3/4/12	51	44
7	3/5/12	46	34
8	3/6/12	44	32
9	3/7/12	48	29
10	3/8/12	60	33
11	3/9/12	49	41
12	3/10/12	45	43
13	3/11/12	44	37
14	3/12/12	47	33
15	3/13/12	42	33
16	3/14/12	46	34

Data Visualization

Temperature Data for Seattle

Date	High (°F)	Low (°F)
3/1/12	43	34
3/2/12	44	39
3/3/12	54	44
3/4/12	51	44
3/5/12	46	34
3/6/12	44	32
3/7/12	48	29
3/8/12	60	33
3/9/12	49	41
3/10/12	45	43
3/11/12	44	37
3/12/12	47	33
3/13/12	42	33
3/14/12	46	34

Chart type in Excel:



Example of a line graph in Excel

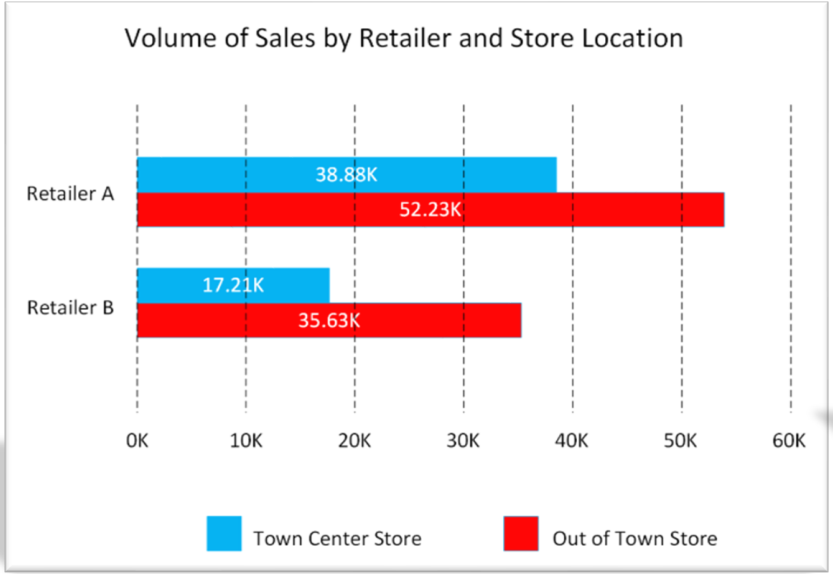
Data Source: http://www.beautifulseattle.com/mthsum.asp

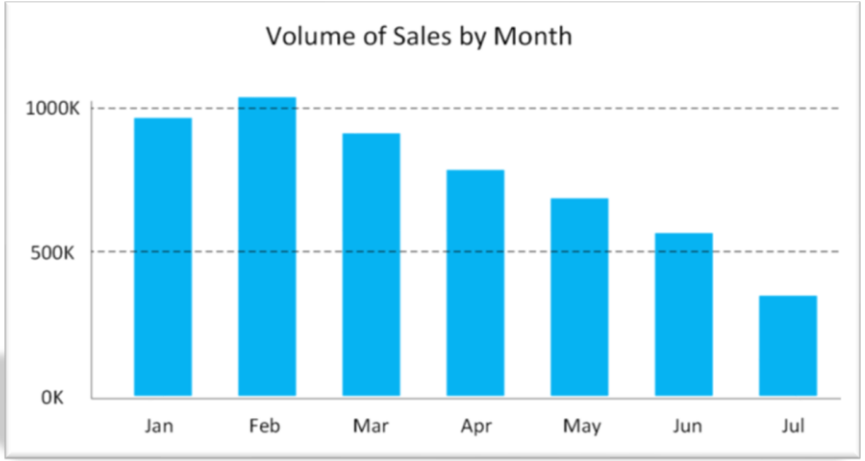
Data Visualization

Most common forms of visualizations are

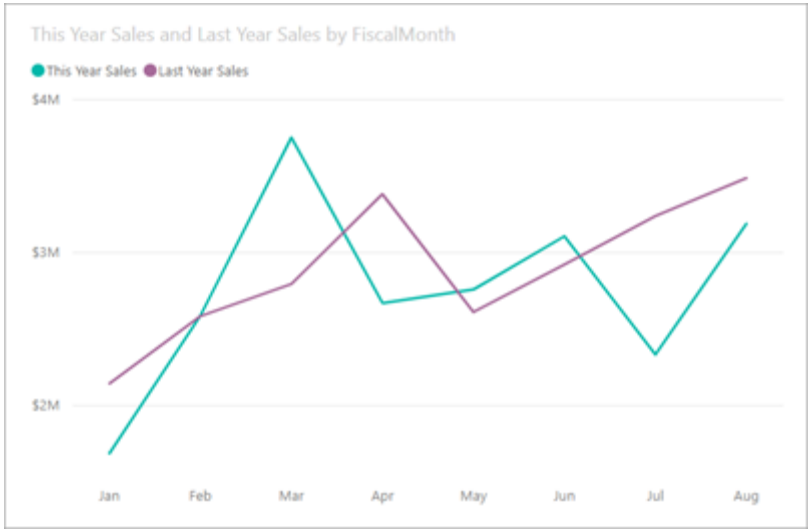
- Bar and column charts
- Line charts
- Pie Charts
- Matrix
- Key Influencers
- Tree map
- Scatter
- Filled Map

Data Visualization – Bar and Column Chart

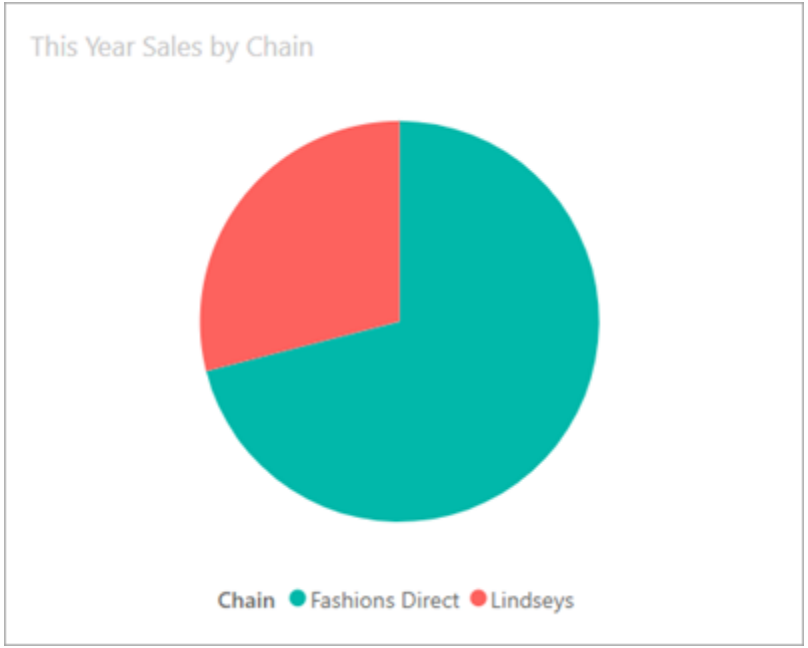




Data Visualization – Line Charts



Data Visualization – Pie Charts

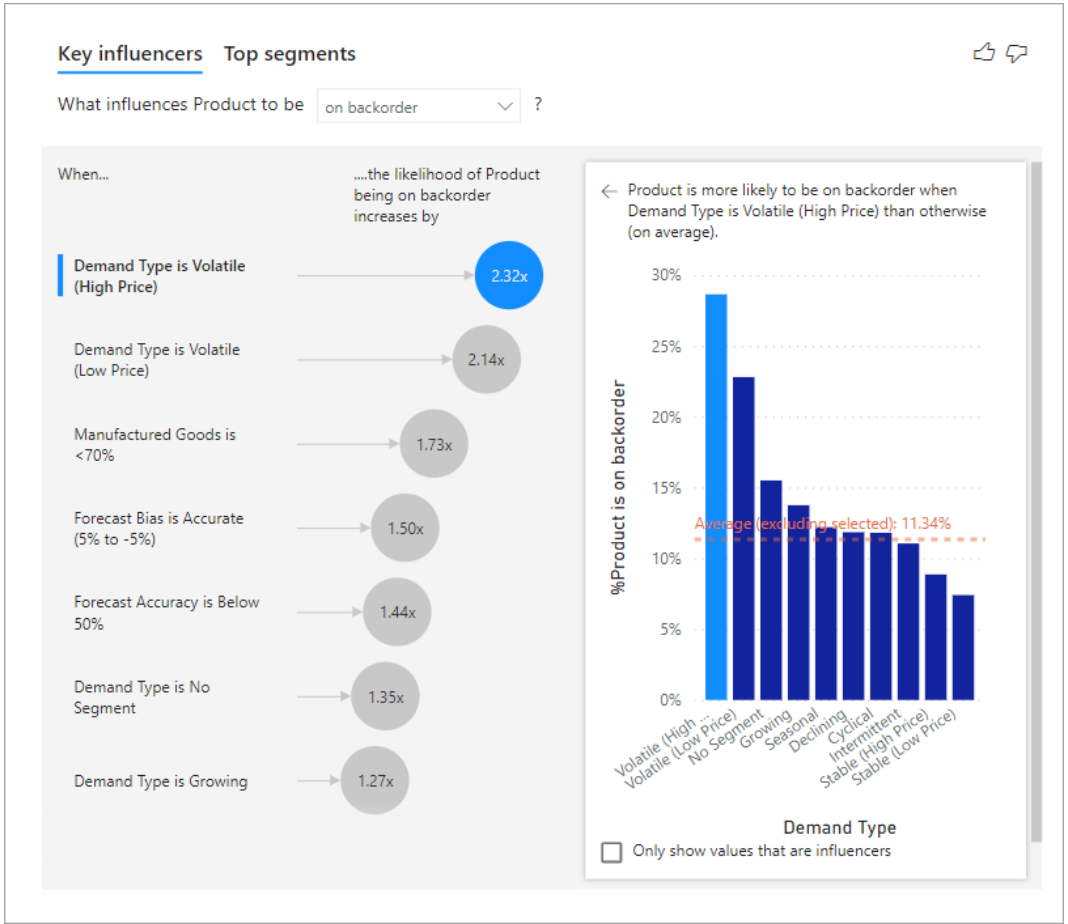


Data Visualization – Matrix

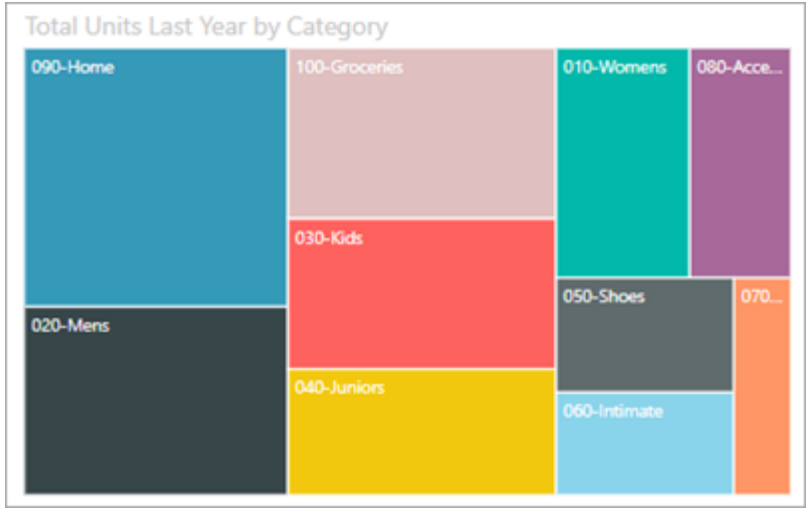
Drill on Rows

Region	Central		East		West		Total	
Sales Stage	Opportunity Count	Revenue	Opportunity Count	Revenue	Opportunity Count	Revenue	**Opportunity Count**	**Revenue**
Lead	102	$507,574,417	114	$473,887,837	52	$256,159,114	**268**	**$1,237,621,368**
Qualify	29	$111,715,461	50	$195,692,154	15	$52,442,363	**94**	**$359,849,978**
Solution	29	$100,743,789	30	$134,347,170	15	$53,441,501	**74**	**$288,532,460**
Proposal	14	$46,722,869	13	$59,970,924	10	$43,032,669	**37**	**$149,726,462**
Finalize	5	$23,302,246	5	$30,696,428	4	$21,176,185	**14**	**$75,174,859**
Total	**179**	**$790,058,782**	**212**	**$894,594,513**	**96**	**$426,251,832**	**487**	**$2,110,905,127**

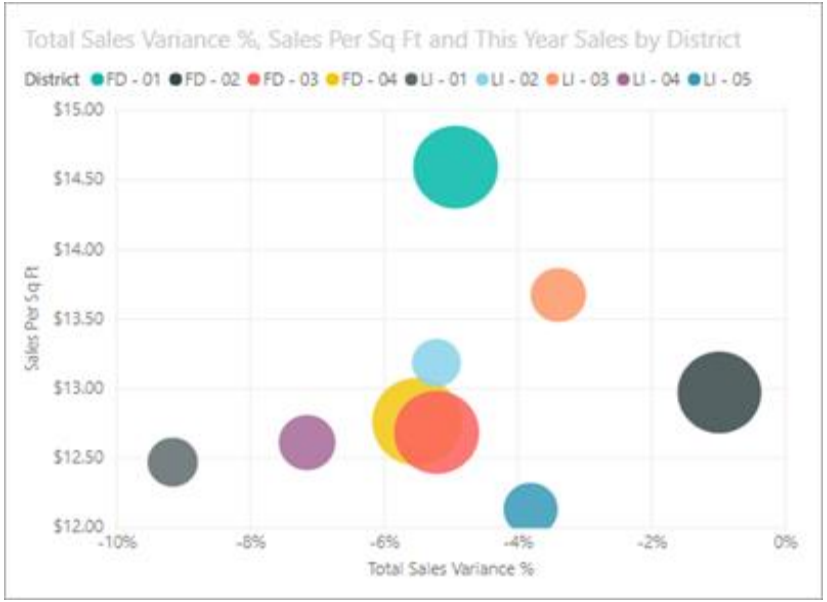
Data Visualization – Key Influencers

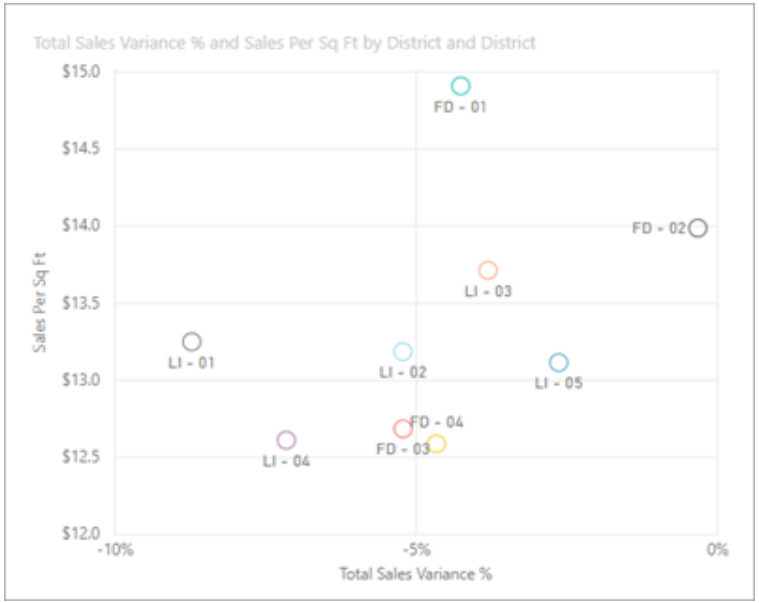


Data Visualization – TreeMaps



Data Visualization – Scatter Charts





sqlbi. OK Viz

THE VISUALS REFERENCE

— FOR MICROSOFT POWER BI —

SEP. 2018

http://sql.bi/visual-reference

COMPARISON

Display measures compared by their magnitude

CHANGE OVER TIME

Display the changing trend of measures

RANKING

Display measures by their rank order

SPATIAL

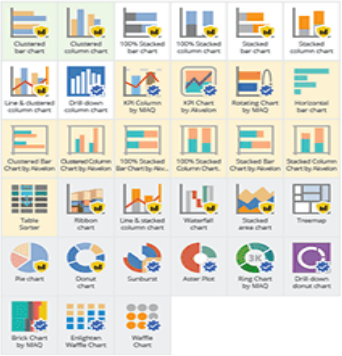
Display measures over spatial maps

FLOW

Display a flow or dynamic relations

PART-TO-WHOLE

Display the parts of a measure



DISTRIBUTION

Display the distribution of a measure

CORRELATION

Display relations between measures

SINGLE

Display single values

FILTER

Control report filters

NARRATIVE

Tell a story with data

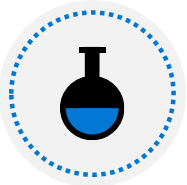
MISCELLANEOUS

Recommended | There is a better alternative | Don't use in the category | Built-in visual | Certified visual | R required

Explore data analytics

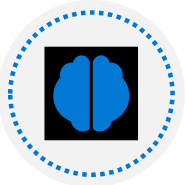
Descriptive

Diagnostic



Predictive

Prescriptive



Cognitive

Descriptive analytics

Descriptive

- Helps us understand - What is happening/happened based on historical data?
- Descriptive Analytics provides us the "**Hindsight**"
- Deals with generating summaries on existing data
- Provides us insights on how everything is going on in the business.
- Doesn't provide any explanation or root cause on why something has happened or happening
- Metrics such as return on investment (ROI), TCO are typical examples of Descriptive Analytics.
- Examples: View of an organization's sales and financial data.

Diagnostic Analytics

Diagnostic

- ***Why*** is it happening in your business? Or “Why things happened”
- Diagnostic Analytics explains the ***root cause*** behind the outcome of descriptive analytics,
- Supplements descriptive analytics outcomes
- Three steps:
    - Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.
    - Collect data that's related to these anomalies.
    - Use statistical techniques to discover relationships and trends that explain these anomalies.

Predictive Analytics

Predictive

❄ ***What's likely to happen*** in the future based on past trends and patterns?

❄ By utilizing various statistical and machine learning algorithms to provide recommendations and provide answers to questions related to what might happen in the future, that cannot be answered by BI

Examples:

❄ Projecting next quarter revenue based on previous quarters and other parameters

❄ Azure Portal predicts your expected bill and usage

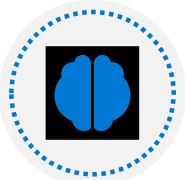
❄ Weather prediction

Prescriptive Analytics

Prescriptive

- Helps you to determine the best course of action to attain/eliminate a certain outcome
- What actions should be taken to achieve a goal or target?
- You can use Prescriptive analytics to advise users on possible outcomes and what should they do to maximize their key business metrics
- Advise on best approach for maximum success
- Google Maps navigation
- Recommendations - If you liked this movie, you might like that one
- Search Engine Optimization tools

Cognitive Analytics



Cognitive

- Combines several intelligent technologies like artificial intelligence, machine-learning algorithms, deep learning etc. to apply human brain like intelligence to perform certain tasks
- Cognitive analytics helps you to learn what might happen if circumstances change, and how you might handle these situations.
- Trained on large real-world datasets and develops a “knowledge” of how a particular task is performed in real world
- Makes predictions based on that “Knowledge”
- Learns and improves with time
- Ex: Analyzing Twitter Tweets to determine brand sentiment

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