Slides

05 October 2021 10:05

dp-900-lecture-

Core Data-Related Azure Services

Azure Storage Accounts
An umbrella service for various storage
Types eg. Table, Files, Blob

Azure Blob Storage
Data which is stored as objects instead of files.
Object storage is distributed storage (spanning multiple machines) for unstructured data

Azure Tables
A key/value NoSQL data store
Intended for simpler projects

Azure Files
A managed file-shared NFS or SMB

Azure Storage Explorer
An application used to explore data within Azure Storage Accounts

Azure Synapse Analytics
Data warehouse and unified ar

CosmoDB
A fully-managed NoSQL databa
Can host various NoSQL engine
Tables, Document, Key/Value,

Azure Data Lake Store (Gen2)
A centralized data repository fo
Blob Storage designed for vasts

Azure Data Analytics
Big Data as a Service (BDaS)
Write U-SQL to return data from

Azure Data Box
Import or export TB of data vi
Into Azure datacenters

Core Data-Related Azure Services

SQL Server on Azure Virtual Machines
When migrating via a lift-and-shift and you want to bring your existing license and need OS access and control of VM

SQL Managed Instances
A managed MS SQL server but with broad adaptably when migrating to Azure

Azure SQL
Fully-managed MS SQL databases.

Azure Databases for <Open-Source>
managed relational databases on Azure
eg. MySQL, PSQL, MariaDB

Azure Cache for REDIS

Microsoft Office 365 Sharel
A shared file-system for org
The company owns all the f
grain role-based access-con

Azure Databricks
A third-party provider part
specializing in Apache Spar
ELT jobs, as well as ML and

Microsoft Power BI
A Business Intelligence too
dashboards and interactive
business decisions

HDInsights
A full-managed Hadoop Syst
many open-source Big Data

In-Memory data-store for returning data
extremely vast but is also extremely volatile

data transformations for Str

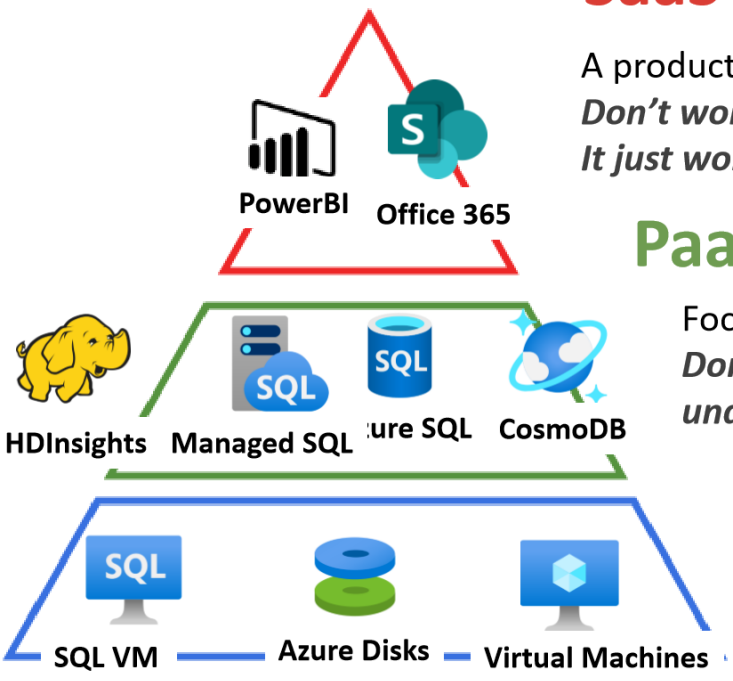
Core Data-Related Azure Services

Azure Data Studio
An IDE that looks very much like Visual Studio Code
But designed around data related tasks. Cross-platform
Similar to SSIS but broader data workloads

Azure Data Factory
A managed ETL/ELT pipeline builder
Easily build transformation pipelines via a web-interface

SQL Server Integration Services (SSIS)
A stand-alone Windows app to prepare data for SQL workloads via transformation pipelines

Types of Cloud Computing



SaaS Software as a Service **For Custon**

A product that is run and managed by the service
Don't worry about how the service is maintained
It just works and remains available.

PaaS Platform as a Service **For Devel**

Focus on the deployment and management
Don't worry about, provisioning, configuri
understanding the hardware or OS.

IaaS Infrastructure as a Service

The basic building blocks for cloud IT.
networking features, computers and
Don't worry about IT staff, data cent

Azure Data Roles

Database Administrator: **configures and maintains a databases** eg. Azure Data services or SQL serve

Responsibilities
- Database management
- Manage security, granting user access
- Backups
- Monitors Performance

Common Tools
Azure Data Studio
SQL Server Management Studio
Azure Portal
Azure CLI

Data Engineer: **Design and implement data tasks** related to the transfer and storage of **big data**

Responsibilities
- Database pipelines and process
- Data ingestion storage
- Prepare data for analytics.
- Prepare data for analytical processing

Common Tools
- Azure Synapse Studio
- SQL
- Azure CLI

Data Analyst: **Analyzes business data** to reveal important information

Responsibilities
- Provides insights into the data
- Visual reporting
- Modeling data for analysis
- Combines data for visualization and analysis

Common Tools
- Power BI Desktop
- Power BI Portal
- Power BI services
- Power BI report bulder

Database Administrator – Common To

configures and maintains a databases

Azure Data Studio Looks like VSC

Connect to Azure SQL, Azure SQL data warehouse, Postgres SQL and SQL Server (big dat
- Various libraries and extensions along with automation tools.
- Graphical interface for managing on-premises and cloud-based data services.
- runs on Windows, macOS, Linux
- Possibly a replacement for SSMS (still lacks some features of SSMS)

SQL Server Management Studio (SSMS)
- Automation tooling for running SQL commands or common database
- Graphical interface for managing on-premises and cloud-based data s
- Runs on Windows
- More mature than Azure Data Studio

Azure Portal and CLI
- Manage SQL database configurations. eg create, deleting, resizing, number of cores
- Manage and provision other Azure Data Services
- Automate the creating, updating or modifying resources via Azure Resource Manage

Data Engineering – Common Tools

Design and implement data tasks related to the transfer and storage of big data

Azure Synapse Studio
azure portal integrated to manage azure synapse, data ingestion (Azure data factory), management of azure synapse assets (SQL Pools/Spark Pool)

Knowledge SQL
Create create databases., tables, views, etc

Azure CLI
Support operations SQL cmd to connect to Microsoft server Azure SQL data and run a talk queries and commands

HDInsights
Streaming data via Apache Kafka or Apache Spark
Applying ELT jobs via HIVE, PIG, Apache Spark

Azure Databricks
Using Apache Spark to create ELT or streaming jobs to data ware houses or data lakes

Data Analyst – Common Tools

Analyzes business data to reveal important information

Power BI Desktop
A stand alone application for data visualization
You can do data modelling
Connect to many data sources
Create interactive reports

Power BI Portal/Power BI Service
A web ui for creating interactive dashboards

Power BI Report builder
Create paginated reports (printable reports)

Data Overview

Data – units of information

Data Documents – types of abstract groupings of data

Data Sets – unstructured logical grouping of data

Data Structures – structured data

Data Types – how single units of data are intended to be used

Batch and Streaming Data
How do we move our data a

Relational and Non Relatio
How do access, query and s

Data Modelling
How do we prepare and

Schemas and Schema
How do we structure

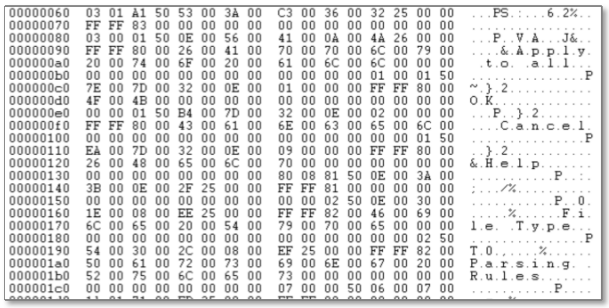
Data Integrity and Data Co
How do we trust our data?

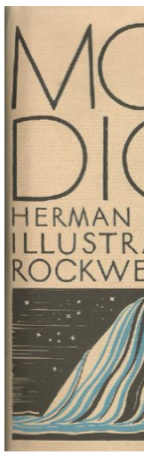
Normalized and Deno
How do we trade qual

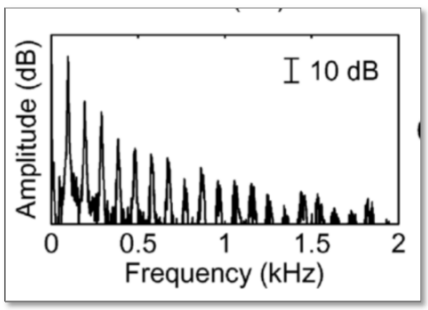
Introduction to Data

What is data?

Data is **units of information** that could be in the form of **numbers**, **text** or **machine code**, **images**, **videos**, **audio** or **physical** *(handwriting)*







$$\int_a^b f'(x)\,dx = f(b) - f(a)$$

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x)$$

Data Documents

What are data documents?

A data document defines the **collective form in which data exists**

A data document defines the collective form in which data exists.

Common types of data documents:

- **Datasets** — a logical grouping of data
- Databases — structured data that can be quickly access and searched
- Datastores — unstructured or semi- structured data to housing data
- Data warehouses — structured or semi-structured data for creating reports ar
- Notebooks — data that is arranged in pages, designed for easy consumption

MNIST Dataset
Dataset

Azure SQL
Database

Azure Data Lake
Datastore

Azure Synapse Analytics
Data warehouse

Jupyt
Digit

Data Sets

What is a dataset?

A data set is a **logical grouping of units of data** that generally are closely related an same data structure.

There are **publicly available data** sets that are used in the **learning of statistics, data analytics, machine learning**

MNIST database

Images of **handwritten digits** used to test classification, clustering, and image processing algorithm.

Commonly used when learning how to build computer vision ML models to translate handwriting into digital text

More Data Sets....

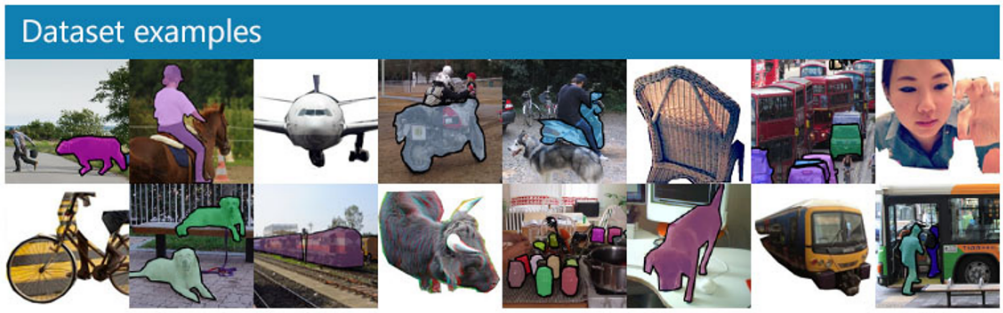
Common Objects In Context (COCO) dataset

A dataset which contains many common images using a JSON file (coco format) that indentify objects or

This dataset fea

- Object segme
- Recognition i
- Superpixel st

segments within an image.

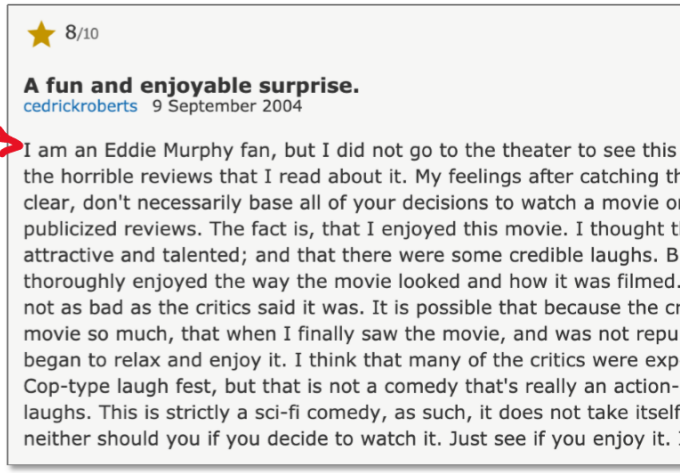


- superpixel st
- 329K images
- 0.5 million ob
- 79 object cat
- 90 stuff categ
- 4 captions pe
- 249,000 peop

More Data Sets...

IMDB Reviews Dataset
A **movie review** dataset with 25,000 highly polar movie reviews for training, and 25,000 for testing

Could be useful for determine customer sentiment analysis.


8/10

A fun and enjoyable surprise.
cedrickroberts 9 September 2004

I am an Eddie Murphy fan, but I did not go to the theater to see this
the horrible reviews that I read about it. My feelings after catching th
clear, don't necessarily base all of your decisions to watch a movie o
publicized reviews. The fact is, that I enjoyed this movie. I thought t
attractive and talented; and that there were some credible laughs. B
thoroughly enjoyed the way the movie looked and how it was filmed.
not as bad as the critics said it was. It is possible that because the cr
movie so much, that when I finally saw the movie, and was not repu
began to relax and enjoy it. I think that many of the critics were exp
Cop-type laugh fest, but that is not a comedy that's really an action-
laughs. This is strictly a sci-fi comedy, as such, it does not take itself
neither should you if you decide to watch it. Just see if you enjoy it.

More Data Sets...

Free Music Archive (FMA)
A dataset for music analysis
- 106,573 tracks
- 163 genres

LibriSpeech
A dataset of 1000 hours of English speech

There are many more datasets online
(some **could be paid**, or you need to **extract the data via an API** or **scrape the dat**

- Crunchbase
- Glassdoor Research
- Open Corporates
- FBI Uniform Crime Reporting
- Uppsala Conflict Data Program
- Dbpedia
- Google Trends
- DataHub – Stock Market
- Center of Disease Control (CDC)
- World Health Organization
- Statista Video Games data
- Data.gov.uk
- Open Data Canada
- NVC Taxi Trip Data
- Weather.gov
- AWS open registry
- Google Public Data
  Reddit Datasets
- USD Food Consumption
- *and many more…*

Data Types

What is a data type?
A data type is ~~a single unit of data~~ that tells a compiler or interpreter (computer **data is intended to be used**.

⚠ The **variety** of data types will **greatly vary based on the computer program**.

Let's us take the most common data types.

Numeric Data Types: A data type involving mathematical **numbers**

- **Integer** – a whole number, (could be negative or positive) -100, 7, 11, 2190381
- **Float** – a number that has a decimal e.g. 1.5 , 0.0, -10.24, 9.432363535345

```
my_int = 1
my_float = 2.2
```

Data Types

Text Data Types: A data type that contains readable and non readable **letters**

- **Character** – a single letter, alphanumeric (A-Z), digit (0-9), blank space, punction, special c
- **String** – a sequence of characters eg. Words, sentences and paragraphs

```
my_char = 'a'
my_string = "We prefer to help ourselves"
```

Composite: A data type that contains **cells of data** that can be **accessed via an index or a key**

- Array – a group of elements that contain the same data type, can be accessed via their index
- Hash (Dictionary) – a group of elements where a key can be used retrieve a value

Composites can be both data-types and data structures

```
my_arr = ['live','long','and','prosper']
my_dict = { "Speed": 1, "Accuracy": 2 }
```

Data Types

Binary Data Type – represented by a **bit or a series of bits (a byte)**, Which is either 0 (off) or 1 (on)

```
one_byte = int('11110000',
```

Boolean Data Type– A datatype that is either **True or False**

- Some languages represents a Boolean as
  - a bit as a Boolean eg. 0 (false) or 1 (true)
  - the first letter eg. t (true) or f (false)

```
my_bool = True
```

Enummeration (Enum) Data Type – a group of constant (unchangeable) variables eg. DIAMOND, SPADE, HEART, CLUBS

- *Can be a data type and/or a data structure, varies on the language*

All cat groups

```
class Shake(Enum):
  VANILLA = 7
  CHOCOLATE = 4
  COOKIES = 9
  MINT = 3

Shake.VANILLA
Shake.CHOCOLATE
Shake.COOKIES
Shake.MINT
```

Schema

What is a schema?

A schema (in terms of databases) is **a formal language which describes the structure of data** (bluepri
A schema can **define many different data structures** that serve different purposes for a database.

Different data structures (relational databases):

- Tables
- Fields
- Relationships
- Views
- Indexes
- Packages
- Procedures
- Queues
- Triggers
- types,
- Sequences
- materialized views
- Synonyms
- database links

Data type

```
ActiveRecord::Schema.define(version: 20140212175753) do

  # These are extensions that must be enabled in order to support
  enable_extension "plpgsql"

  create_table "activities", force: true do |t|
    t.integer  "company_id"
    t.integer  "project_id"
    t.integer  "target_id"
    t.string   "target_type"
    t.integer  "user_id"
    t.string   "name"
    t.string   "key"
    t.datetime "created_at",  null: false
    t.datetime "updated_at",  null: false
```

- Functions
- XML schemas
- Directories

↑ use schema to create these

```
  t.json     "data"
end

add_index "activities", ["company_id"], name: "index_activities_
add_index "activities", ["project_id"], name: "index_activities_
add_index "activities", ["target_id"], name: "index_activities_o
  t.integer  "project_id"
end
# ...
```

A Ruby on Rails schema that defines the structure for

Schemaless

What is a schemaless?
Schemaless is when the **primary "cell" of database can accept many types**
This allow developers to **forgo upfront data modelling**

Common schemaless databases are:
- Key/Value
- Document
- Columns
  - Wide Column
- Graph

Data type more flexible

Query and Querying

SELECT * FROM CrewMembers

What is a query?
A query is a request for data results (reads) or to perform operatio
updating deleting data (writes).
A query can perform maintenance operations on the data and is n
just working with the data that resides within the database.

Id	Name	Title	StarDate
10	Picard	Captain	41133.7
11	Riker	Commander	41154.2
12	Data	Lt Commander	41114.3

What is a data result?
Results are the data returned from a qu

What is querying
That act of performing a query

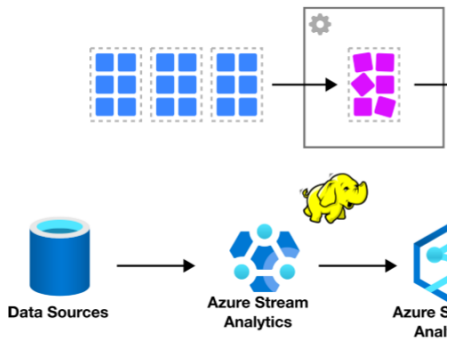
What is query language?
A scripting or programming language d
format to submit a request or actions t
Notable query languages:
- SQL

13	Troi	Lt Commander	41412.1
14	Crusher	Command	41520.4

- GraphSQL
- Kusto
- Xpath
- Gremlin

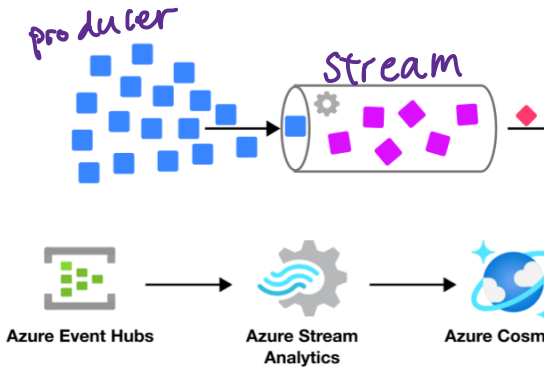
Batch vs Stream Processing

Batch Processing
When you send batches (a collection) of data to be processed.
Batches are generally scheduled: eg. Every day at 1PM
Batches are not real-time
Batches processing is ideal for very large processing workloads
Batch processing is more cost-effective



Stream Processing
When you process data as soon as it arrives:
- **Produces** will send data to a stream and
- **Consumers** will pull from the stream

Stream processing is good for real-time analytics or real-time processing (streaming video)
Much more expensive than batch processing



Relational Data

Tables
A logical grouping of rows and columns. Think like a Excel spreadsheet
- Tabular data – data that makes use of table data structures

Views
Views is a result set of a stored query on data **stored in memory** (a temporary or virtual table)

Materialized Views
Material Views is a result set of stored query on data **stored on disk**

Indexes
A copy of your data sorted by one or multiple columns for faster reads at the cost of storage

Constraints
rules applied to writes, that can ensure data integrity: eg. don't allow duplicate records

Triggers
a function that is trigger on specific database events

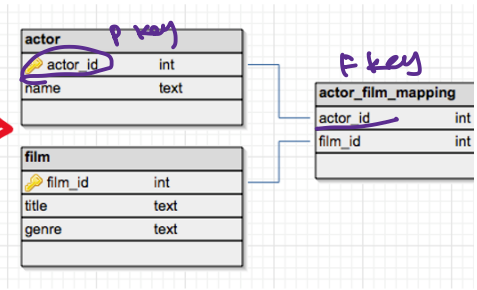
Primary Key
one or multiple columns that uniquely identify a table in a row eg. Id

Forgien key
a column which holds the value of primary key from another key to establish a relationship

- A relationship is when two tables have a reference to one another to join data together

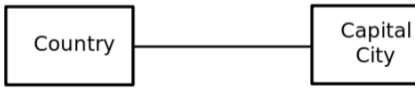
Relational Data – Relationships

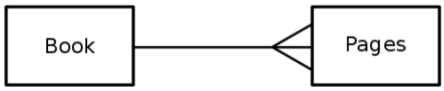
Relational databases **establish relationships to other tables** via foreign keys Referencing another table's primary key.



These are the relationships found in a relational data:

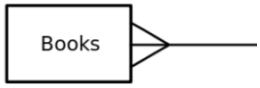
One-to-one
A monkey has a banana





One-to-many
A store has many customers

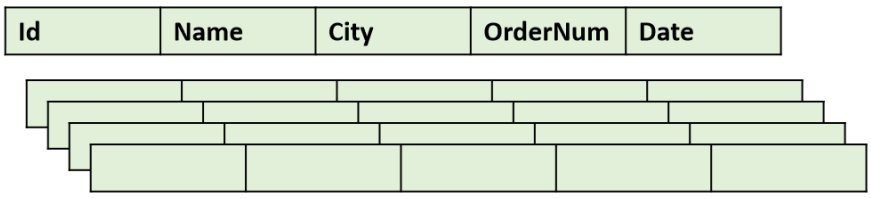
Many-to-many
A project has many tasks and
Tasks can belong to many projects



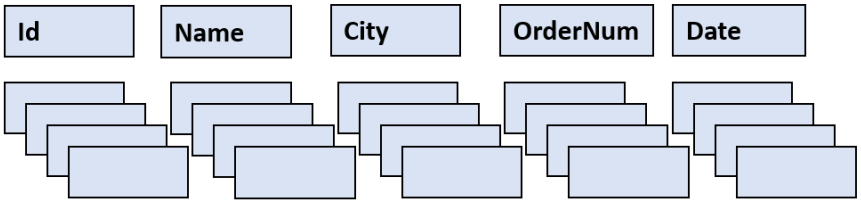
Many to Many (via Join/Junction Table)
A student has many classes through enrollments
A class has many students through enrollments

Row-store vs Column-store

When looking at tabular data there are two ways data can be stored:
Row-store or Column Store



Row-store
- Data is organized in rows
- Traditional relational databases are
- Good for general purpose database
- Suited for Online transaction proces
- Great when needing all possible col
- Not the best at analytics or massive



Column-store
- Data is organized into columns
  - Faster at aggerating values f
- NoSQL store or SQL-Like database
- Great for vast amount of data
- Suited for Online analytical proces
- Great when you only need a few c

Indexes

A database index is a data structure that improves the speed of reads from the
by **storing the same or partial redundant data** organized in a more efficient logi

The logical ordering is commonly determined by one ore more columns: sort key(s)

A common data structure of an index is a **Balanced Tree (B-Tree)**

Table (phone book) **Inc**

Name	Number	ID
Andrew	626-9009	1
Carly	641-5324	2
Rishab	345-4121	3
Cindy	767-3423	4
Peter	623-2413	5
Lisa	767-5235	6
Otto	344-5353	7
Mona	345-6189	8
Zack	626-4421	9
Maya	767-7771	10

Creating an index in Postgres

```
CREATE INDEX idx_address_phone
ON address(phone)
```

Data Integrity and Data Corruption

Data integrity is the **maintenance and assurance of, data accuracy and consistency** over

Used at as **proxy term for data quality**, data validation is a pre-requisite for data integ

The goal of data integrity:
ensure data is recorded exactly as intended

Data integrity is the ***opposite*** of Data corruption

Data corruption is the **act or state of data not being in the intended state** and will result in data loss or misinformation
Data corruption occurs when untended changes result when reading and writing:

- Unexpected hardware failure
- Human error when inputting, modifying data
- Malicious actors with intent of corrupting your data
- Unforeseen side affects for automated operations via computer code

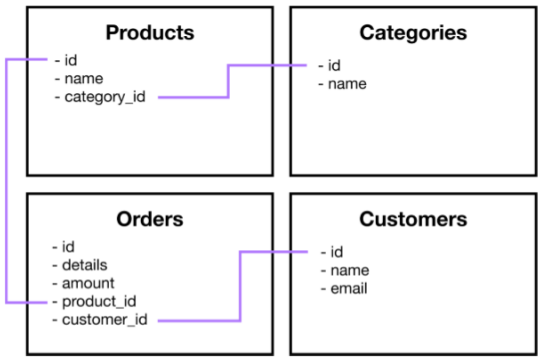
How do we ensure data integrity?

- Have a well defined and documer
- Logical constraints on your databa
- Redundant and versions of your d restore
- Human analysis of the data
- Hash functions to determine if ch tampered
- Principle of least-privilege, (limitir actions for specific user roles)

Normalized vs Denormalized Data

Relational d
denormalise

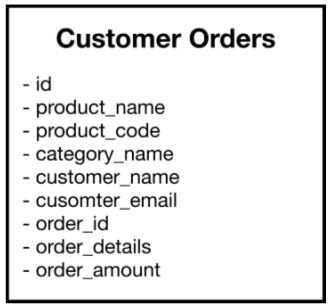
Normalized
A schema design to store **non-redundant** and **consistent data**



- Data Integrity is maintained
- Little to no redundant data
- Many tables
- Optimizes for storage of data

Denormalized
A schema that **combines data** so tha **accessing data (querying) is fast**

combine

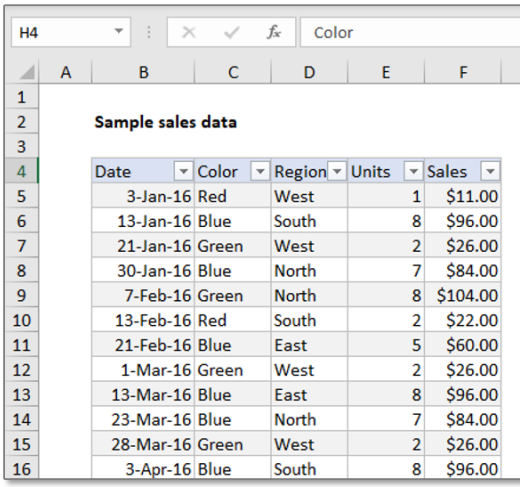


- Data Integrity is not maintained
- Redundant data is common
- Fewer tables
- Excessive data, storage is less opt

Pivot Table

A pivot table is **a table of statistics that summarizes** the data of a more exte table from a: Database, Spreadsheet or Business intelligence (BI) tool

- Pivot tables are a technique in **data processing**
- They arrange and rearrange (or "pivot") statistics in order to **draw attention to useful information**
- This leads to **finding figures and facts quickly** making them integral to data analysis.



H4 | Color

	A	B	C	D	E	F
1						
2		**Sample sales data**				
3						
4		Date	Color	Region	Units	Sales
5		3-Jan-16	Red	West	1	$11.00
6		13-Jan-16	Blue	South	8	$96.00
7		21-Jan-16	Green	West	2	$26.00
8		30-Jan-16	Blue	North	7	$84.00
9		7-Feb-16	Green	North	8	$104.00
10		13-Feb-16	Red	South	2	$22.00
11		21-Feb-16	Blue	East	5	$60.00
12		1-Mar-16	Green	West	2	$26.00
13		13-Mar-16	Blue	East	8	$96.00
14		23-Mar-16	Blue	North	7	$84.00
15		28-Mar-16	Green	West	2	$26.00
16		3-Apr-16	Blue	South	8	$96.00

https://exceljet.

In **Micorsoft Excel** its very easy to create Pivot Tables. Think of a pivot table as an interactive report where you can quickly aggerate (group) your data based on various factors eg.

- By Year Month, Week or Day
- Sum, Average, Mix or Max

"PivotTable" used to be a trademarked word owned by Microsoft

Strongly Consistent vs Eventually Consis

What is data consistency?

When data being kept in two different place and **whether the data exactly match** or do not ma

When you have to have duplicates of your data in many places and need to keep them up-to-date to be exact matching, based on how data is transmitted and service levels cloud service providers will use these two terms:

Strongly Consistent
Every time you request data (query) you can expect consistent data to be returned with x time (1 seconds)

We will never return to you old data. But you will have to wait at least 2 seconds for the query to return

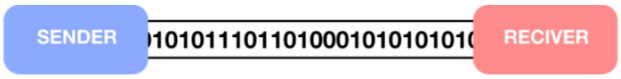
Eventually Consistent
When you request data you may get ba
within 2 seconds.

We are giving you whatever data is currentl
may get new data or old data, but if you wa
generally be up to date.

Synchronous vs Asynchronous

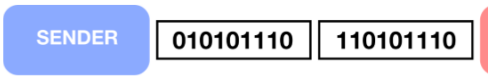
Synchronous and Asynchronous can refer to **mechanism for data transmission or data replicat**

Synchronous
continuous stream of data that is synchronized by a timer or clock (guarantee of time)
Can only access data once transfer is complete
- Guaranteed consistency of data return at time of access
- Slower access times



A company has a primary database, but they need to have a backup database in case their primary database fails. The company cannot lose any data, so everything must be in-sync. The database is not going to be accessed while it is standing by to act as replacement.

Asynchronous
continuous stream of data separated b
(no guarantee of time)
Can access data anytime but may return old
placeholder
- Faster access times, not guarantee of co

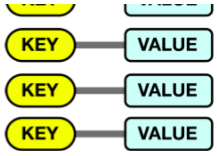


A company has a primary database,
a read-replica (copy of the database)
analytics person can create computa
intensive reports that do not impact
database. It does not matter if the d
1-to-1 at time of access

Non Relational Data

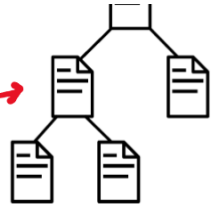
A non-relational database **stores data in a non-tabular form** and will be optimized for different kind

Types of non-relational databases:
Key/Value

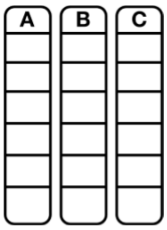


- Each value has a key
- Designed to scale
- Only simple lookups

Document



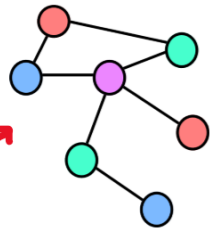
- Primary entity is a JSON-like data-structure called a document



Columnar

- Has a table-like structure but data is stored around columns instead of rows

Graph



- Data is represented with nodes and structures. Where relationships matter

Sometimes non-relational database can be both Key/Value and Document
eg. Azure Cosmo DB or Amazon DynamoDB

Wide-Column vs Columnar

Columnar data store will store each column separately on disk

Wide-column database is a type of colu
database that supports a column family
together on disk, not just a single colum

Data Sources

What is a data source?

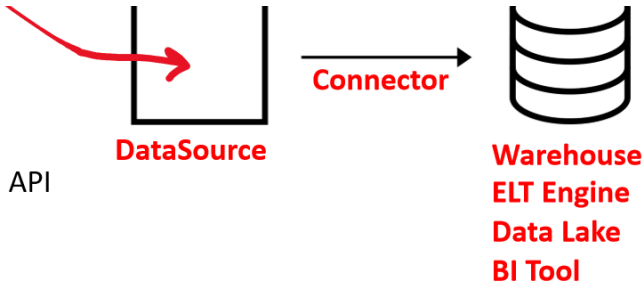
A data source is **where data originates from**.

An analytics tool may be connected to various data sources to create a visualizati

A **data source** could be a:

Data fo

- Data lake
- Data warehouse
- Datastore
- Database
- Data requested on demand from an API endpoint from a web-app
- Flat files (e.g. exel spreadsheet)

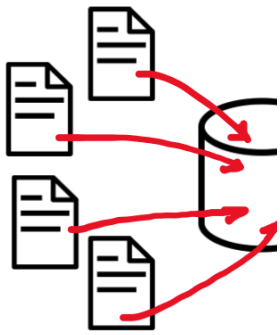


Extracting data from data sources
A data tool like Business Intelligence (BI) software would establish a connection to multiple data sources. A BI will *extract* data which could be could pull data at the time report, or could pull dat schedule, or data could be streamed. The mechanism for extracting data will vary per data sourc

What is Datastore?

A Datastore a **repository for persistently storing and managing collections** of **unstructured or semi-structured data**

Data store is a very broad term, and interchangeable used with databases. But generally a data store indicates working unstructured or semi-structured data



A data store can be specialized in storing:

- Flat Files (documents, spreadsheets)
- Email (an email server)
- Databases (complex data stores developed using formal design and modeling techniques)
  - Relational Databases
  - NoSQL Databases
  - Object-Oriented Databases
- Data stores designed to be distributed across many machines
- Directory services

What is a Database?

A database is a **data-store that stores semi-structured and structured data**.

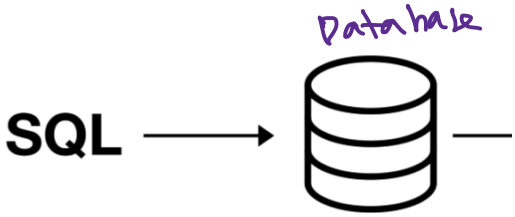
A database is more **complex data stores** because it **requires using formal des modeling techniques**

Databases can be generally categorized as either:

- **Relational databases**
  - Structured data that strongly represents tabular data (tables, rows and columns)
    - Row-oriented or Columnar-oriented

- **Non-relational databases**
  - Semi-structured that may or may not distantly resemble tabular data.

Databases have a rich set of functionality: eg. SQL
- specialized language to query (retrieve data)
- specialized modeling strategies to optimize retrieval for different use cases
- more fine tune control over the transformation of the data into useful data structures or reports



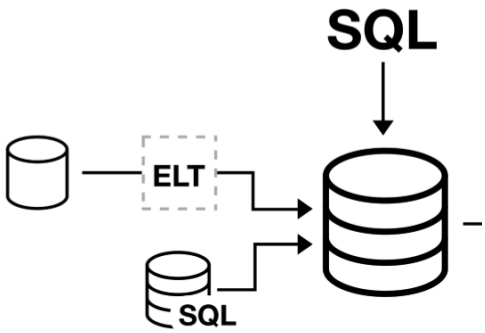
Normally a databases infers someone is using a **relational row-oriented data st**

What is Data Warehouse?

A relational datastore designed for **analytic workloads**, which is generally **column-o**

Companies will have **terabytes and millions of rows of data**, and they need a fast way to be able to produce analytics reports

Data warehouses generally perform **aggregation**
- aggregation is grouping data eg. find a total or average
- Data warehouses are optimized around columns since they need to quickly aggerate column data



Data warehouses are generally designed be HOT
- Hot means they can returned queries very very fast even though they have vast amounts of data

Data warehouses are infrequently accessed meaning they aren't intended for real-time reporting but maybe once or twice a a day or once a week to generate business and user reports.

generally are not for tra

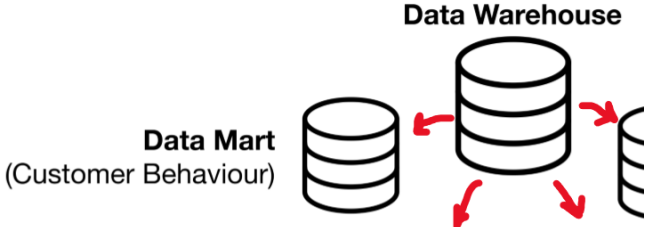
A data warehouse needs to consume data from a relational databases on a regular basis.

What is Data Mart?

A Data mart **subset of a data warehouse**
A data mart will store under **100 GB and has a single business focus**

Data mart allows different teams or departments to have control over their own dataset for their specific use case

Data marts are generally designed be ***read-only.***



Smaller dataset, lower cost of query

Data marts also increase the frequency at which data can be accessed.

The cost to query the data is much lower and so queries can be performed multiple times a day or even hourly.

What is a Data Lake?

A **data lake** is **a centralized storage repository that holds a vast amount of raw data** (big either a semi-structured or unstructured format.

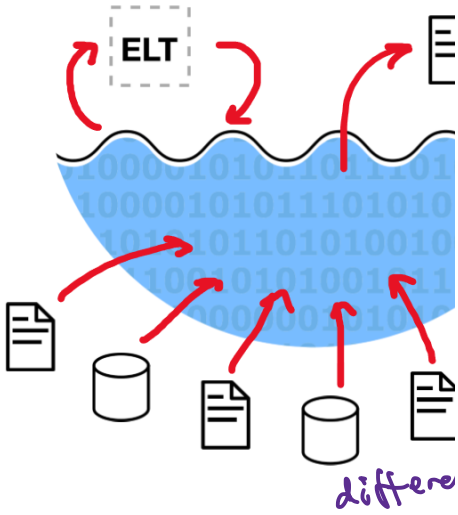
A data lake lets you store all your data without careful design or having to answer questions on the future use of the data. ***(Hording for data scientist)***

A data lake is commonly accessed for data workloads such as:

- Visualizations (Business Intelligence)
- Real-time analytics
- Machine Learning
- On-premise data

Data lakes are great for data-scientists but its very hard to use data lake for BI reporting

If data lakes are not well-maintained they can become **data-swamps** (a mess of data)



What is a Data Lakehouse?

A Data Lakehouse combines the best elements of a **data lake** and a **data ware**

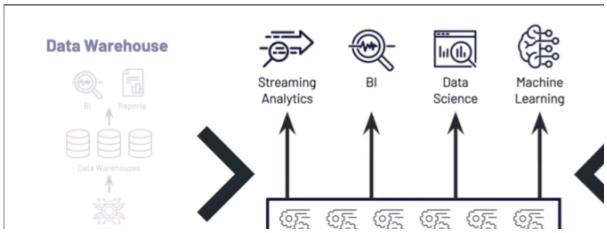
Data Lakehouses compared to a Data warehouse can:

- support video, audio and text files
- Support data science and ML workloads
- have support for both streaming and ELT
- Work with many open-source formats
- Data will generally reside in a data lake or blob stores

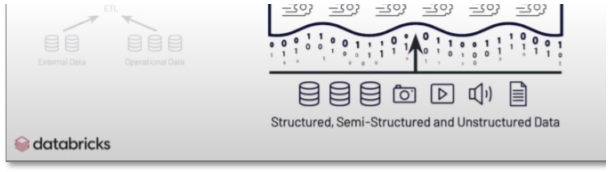
Data Lakehouses compared to a Data la

- perform BI tasks very well
- much easier to setup and maintain
- has management features to avoid a data swamp
- more performant than a data lake

DELTA LAKE



An example of a Data Lakehouse is **Apache Delta Lake** →



Data Sources Comparisons

Datastore — a repository for data for unstructured or semi-structured data

Database — a repository for structured data, commonly refers to Relational databases

Data warehouse — a datastore designed for analytic workloads, generally columnar store

Data Lake — a centralized datastore for store vast amounts of data with distrubted storage, genera for data analytics

Data Lakehouse — combines the best elements of a data lake and data warehouse

Data mart — a subset of data warehorse, it stores under 100 GB and has a single business focus

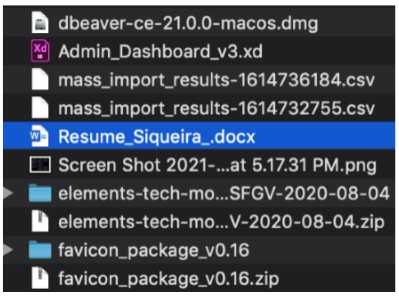
Data Structures

What is a Data Structure?

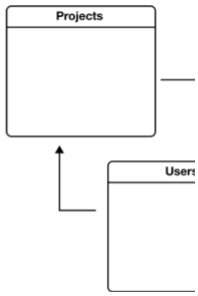
Data that is organized in a **specific storage format**, that enable easy access and mo
A data structure can store various data types.

Data can be abstractly described to have a degree of structure:

- **Unstructured** — a bunch of lose data that has no organization or possibl
- **Semi-Structured** — data that can be browsed or searched (with limitatic
- **Structured** — data that can be easily browsed or searched



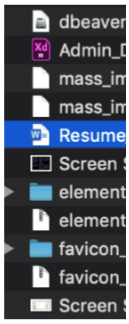

```
<?xml version="1.0"?>
<Root>
  <Customer>
    <Order OrderId="152313">
      <Name>Self-sealing stem bolts</Name>
      <Quantity>100</Quantity>
    </Order>
  </Customer>
</Root>
```



Screen Shot 2021-...at 1.42.23 PM.png

Unstructured Semi-structured Struct

Data Structures – Unstructured



Unstructured data is just a **bunch of loose data**, think of a junk folder on your computer with a bunch of random files, not optimized for search or analysis, or simply no relation between various data.

Microsoft and Azure services that store unstructured data

Microsoft SharePoint
Shared documents for an organization

Azure Blob Storage
Unstructured object datastore

Azure Files
Mountable file system for storing unstructured files

Azure D
For big dat
From many

Data Structures – Semi-Structured

Semi-structured data is (*no schema) data **has some form of relationship**, its easy to browse data to find related data, you can search data but there are limitations or when you search you will pay at a computative or operational cost

Concreate semi-structure data structures:
- XML, JSON, AVRO, PARQUET

Azure and other services that store semi-structured data

Azure Tables
A key/value data store

Azure Cosmos DB
A document data store designed for global scale

MongoDB
Open-source document store NoSQL database

Apach
Open-source
NoSQ

Semi Structured Data structure

What is semi-structured data?
Semi-structured data is data **that contains field**s.
The fields don't have to be the same in every entity.
You only define the fields that you need on a per-entity basis.

Common semi-structured data structures:

JavaScript Object Notation (JSON)
Format used in JavaScript notation; Store data in memory, read and write from files.

Apache Optimized Row Columnar format (ORC)
organizes data into columns rather than rows (columnar store data structure).

Apache Parquet
Another columnar data structure. A Parquet file contains row groups.

Apache AVRO
row-based format. Each record contains a header that describes the structure of the data in the reco

Semi Structured – JSON

JSON (JavaScript Object Notation) is a lightweight data-interchange form
- It is easy for humans to read and write.
- It is easy for machines to parse and generate
- It is based on a subset of the JavaScript

```
{
  "starships": {
    "enterprise": {
      registry: [
        "NCC-1701",
        "NCC-1701-B",
        "NCC-1701-C",
        "NCC-1701-D"
      ]
    }
  }
}
```

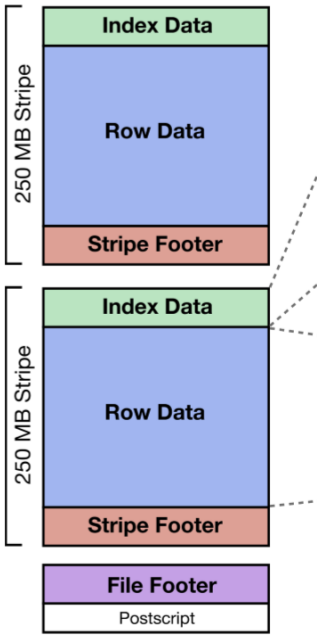
JSON is built on two structures
1. **A collection of name/value pairs**
   - In other languages: realized as an *object*, re dictionary, hash table, keyed list, or associa
2. **An ordered list of values**
   - In other languages: realized as an *array*, ve sequence

JSON is a **text format** that is completely language independent

Semi Structured – ORC

Apache ORC (Optimized Row Columnar) a storage format of the Apache Hadoop

- It is similar to RCFile and Parquet files and is the successor to RCFiles
- It was developed by Facebook to supports columnar reads, predictive pushdown and lazy reads
- It is more storage efficient than RCFiles (taking up 75% less space)
- ORC only supports Hadoop's HIVE and PIG
- ORC performs better with HIVE than Parquet files
- ORC files are organized into **stripes of data**



The Anatomy of an ORC file

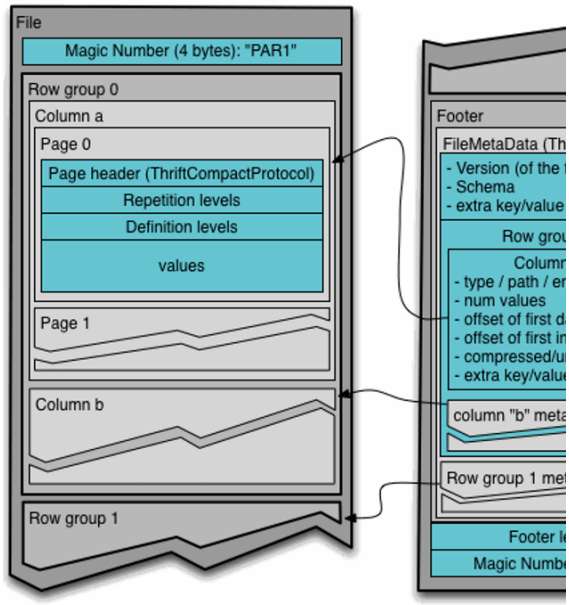
- **File footer** stores auxiliary information
  - list of stripes in the file
  - number of rows per stripe
  - each column's data type.
  - column-level aggregates count, min, max, and sum
- **stripe footer** contains a directory of stream locations
- **Row data** is used in table scans
- **Index data** includes min and max values for each column and the row positions within each column
- The default stripe size is **250 MB**
- Large stripe sizes enable large, efficient reads from HDFS

Hadoop's file system

Semi Structured – Parquet

Apache Parquet is a **columnar storage file format** available to **any project in the Hadoop ecosystem** (Hive, Hbase, MapReduce, Pig, Spark)

- Parquet is built to support very efficient compression and encoding scheme
- uses the record shredding and assembly algorithm

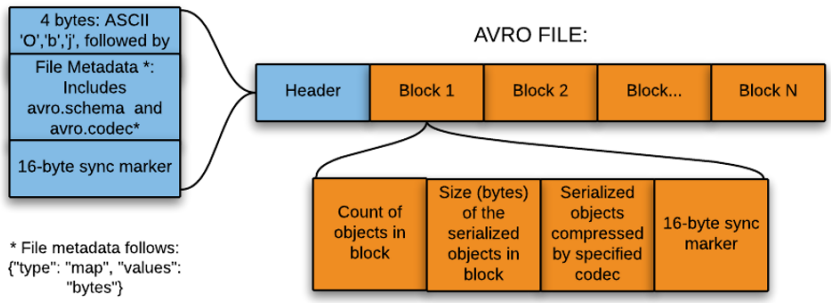


Semi Structured – AVRO

Apache AVRO is a row-based format that provides:

- Rich data structures.
- A compact, fast, binary data format.
- A container file, to store persistent data.
- Remote procedure call (RPC).
- Simple integration with dynamic languages

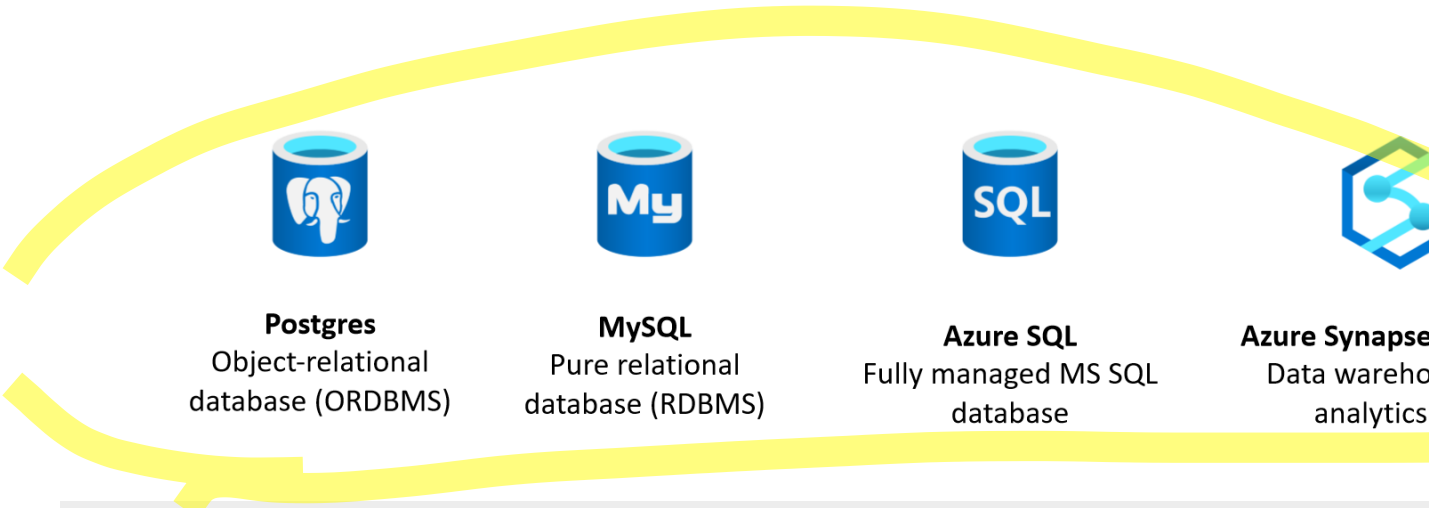
Avro provides functionality similar to systems such as Thrift, Protocol Buffers



Data Structures – Structured

Structured data is (schema) data has a relationship, its easy to browse to find related data, its easy to search data. The most common structured data is tabular data (representing row and columns)

The most common data structure is **tabular data**



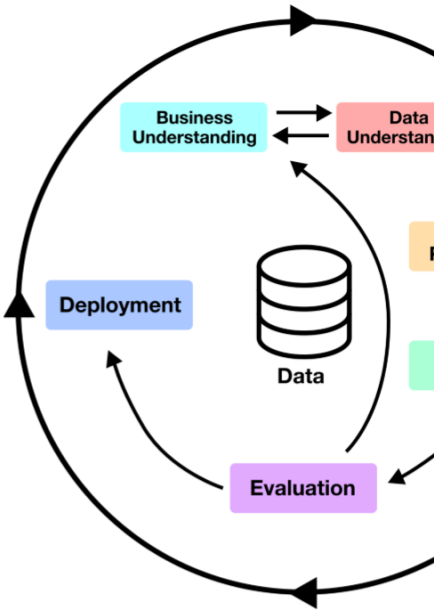
What is Data Mining?

What is Data Mining?

The **extraction of patterns and knowledge** from large amounts of data **(not the extractio**

Cross-industry standard process for data mining (CRISP-DM) defines Data Mining into 6 phases:

1. Business understanding
What does the business need
2. Data understanding
What data do we have, and what data do we need?
3. Data preparation
How do we organize the data for modeling?
4. Modeling
What modeling techniques should we apply?
5. Evaluation
Which data model best meets the business objective?
6. Deployment
How do people access the data?



Data Mining Methods

Data Mining methods or techniques is a way to **find valid patterns and relationships**

Classification
classify data in different classes
Clustering
a division of information into groups of connected objects
Regression
identify and analyze the relationship between variables because of the presence of the other factor
Sequential
evaluating sequential data to discover sequential patterns
Association Rules
discover a link between two or more items, finds a hidden pattern in the data set
These common ***constraints*** (math formulas) are used to determine significant and interesting links:

- Support — indication of how frequently the itemset appears in the dataset.
- Confidence — indication of how often the rule has been found to be true
- Lift — indication of importance compared to other items
- Conviction — indication of the strength of the rule from statistical independence

Outer Detection outliers
observation of data items in the data set, which do not match an expected pattern or expected beha
Prediction
used a combination of other data mining techniques such as trends, clustering, classification to pred

What is Data Wrangling?

What is Data Wrangling?
The **process of transforming and mapping data from one "raw" data form into another f**
intent of making it more appropriate and valuable for a variety of downstream purposes s
Also known as data munging

There **6 core steps** behind data wrangling:

There **6 core steps** behind data wrangling:

1] Discovery
understand what your data is about and keep in mind domain specific details about your data as you move through the other steps

2] Structuring
you need to organize your content into that will be easier to work for our end

3] Cleaning
remove outliers, change null values, remove duplicates, remove special characters, standardize formatting

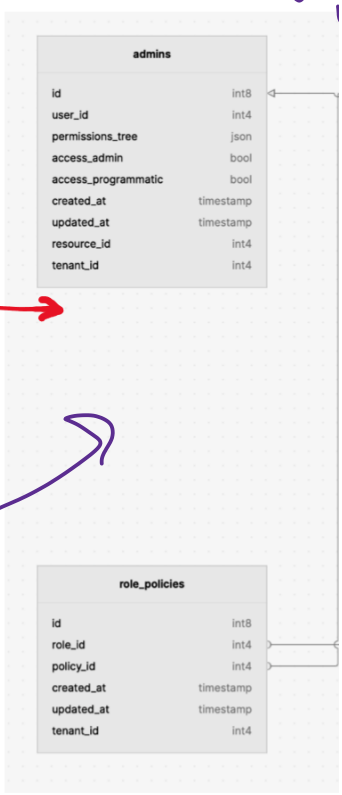
4] Enriching
appending or enhancing collected data context obtained from additional sour

5] Validating
authenticate the reliability, quality, and safety of the data

6] Publishing
place your data in a datastore so you c

What is Data Modeling?

What is a Data Model?
an abstract model that **organizes elements of data and standardizes how they relate to one another** and to the properties of real-world entities eg. A data model could be a **relational database** that contains many tables.

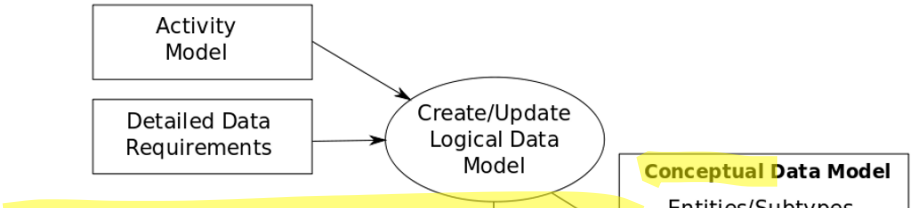


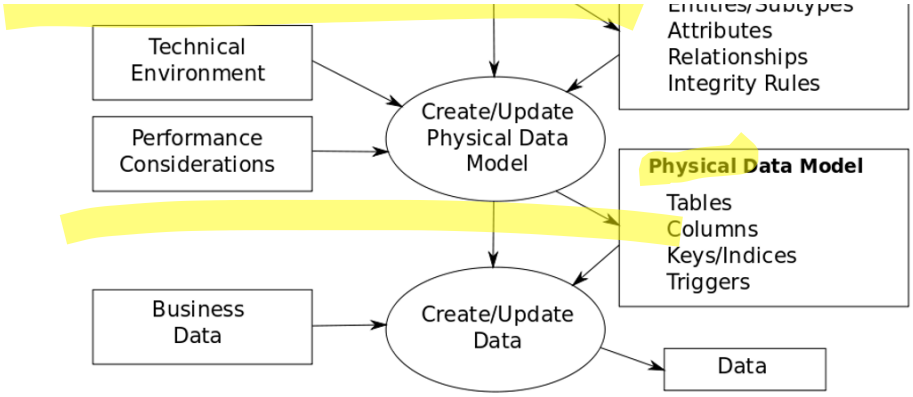
A data model could be:

- **Conceptual**
  - How data is represented at the organization level abstractly without concretely defining how it work with in software
    - eg. People, Orders, Projects, Relationships
- **Logical**
  - How data is presented in software
    - eg. tables and columns, object oriented classes
- **Physical**
  - How data is physical stored:
    - eg. partitions, CPUs, tablespaces

What is Data Modeling?

What is data modeling?
A process used to **define and analyze data requirements needed to support the business processe** scope of corresponding information systems in organizations



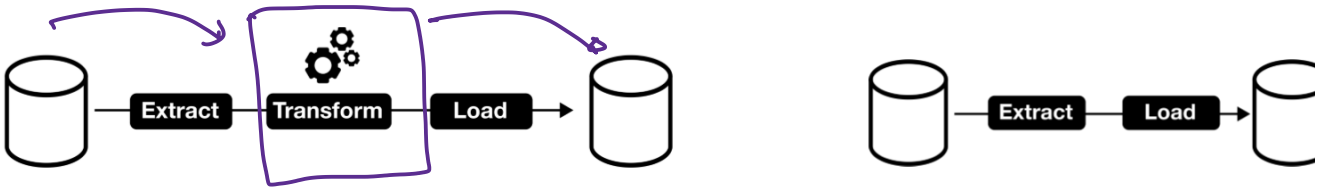


ETL vs ELT

ETL and ELT is used when you want to **move data from one location to another**, where the datasto have different data structures so you need to **transform the data for the target system**

relational No SQL

Load to an intermediate VM eg. Moving SQL Server database into CosmoDB Tables



Extract, Transform and Load (ETL)

- loads data first into a staging server and then into the target system
- used for on-premises, relational and structured data
- used for a small amount of data
- doesn't provide data lake support
- easy to implement
- Mostly supports relational data

Extract, Load and Transform (ELT)

- loads data directly into the target system
- used for scalable cloud structured and un
- used for large amounts of data
- provides data lake support.
- requires specialized skills to implement a
- Support for unstructured data readily ava

Data Analytics

What is Data Analytics?

Data analytics is concerned with **examining, transforming, and arrangin data** so that you can **extract and study useful information**.

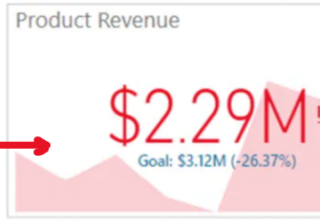
A data analyst commonly uses SQL, Business Intelligence (BI) tools and Spreadsheets

Data Ingestion | Data Cleaning and | Dimensionality | Data Analysis

Data Analytics Workflow

Key Performance Indicators (KPIs)

Key Performance Indicators (KPIs) are type of performance measurement that a company or organization to determine **performance over time**



KPIs can **evaluate the success** of an organization or of a specific organization a

There are two categories of measurements for KPIs

1. Quantitative
- properties can be measured with a numerical result
- facts presented with a specific value
  - eg. monthly revenue, number of signups, number of reported defects

2. Qualitative
- properties that are observed and can generally not be measured with a numerica
- numeric or textual value that represent of personal feelings, tastes, or opinions
  - eg. Customer sentiment

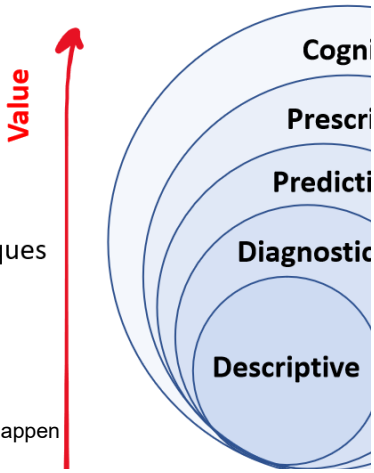
Data Analytic Techniques

Descriptive Analytics — What happened?
- Specialized metrics
  - Key Performance Indicators (KPI)
  - Return on Investment (ROI)
- Generating sales and financial reports
- Accurate, comprehensive , live-data and effective visualizations

Diagnostic Analytics — Why did it happen?
- Supplemental to descriptive analytics
- Drill down, or investigate descriptive metrics to determine root cause
- Find and isolate anomalies into its own datasets and apply statistical techniques

Predictive Analytics — What will happen?
- Use historical data to predict trends or reoccurrence
- Statistical and Machine Learning techniques applied
  - eg. Neural Networks, Decision Trees, Regression, Classification

Predictive Analytics — What will happen? Prescriptive analytics - how can we make it happen
- Goes a step further than predictive and uses Machine Learning by ingesting



- Goes a step further than predictive and uses Machine Learning by ingesting hybrid data to predict future scenarios that are exploitable

Cognitive Analytics — What-if this happens?

- Using analytics to draw patterns to create what-if scenarios and what actions can be taken if those scenario bcome reality
- Uses ML, Natural Language Processing (NLP)
  - Eg call center conversation logs and product reviews

Microsoft OneDrive

Microsoft OneDrive is **a storage and storage synchronization service** for files which r
Similar products: DropBox, GoogleDrive, Box and *Microsoft SharePoint

OneDrive is intended for personal storage for a single individual
You pay for different sizes of storage (5GB ***free***, 100GB, 1TB, 6TB)
You do not worry about the underlying hardware's the, durability, resilience, fault tolerance, or av

Files can be shared easily to other users via:

- a shareable link
- or to a specific email that also has a OneDrive Account

Files are accessed via:

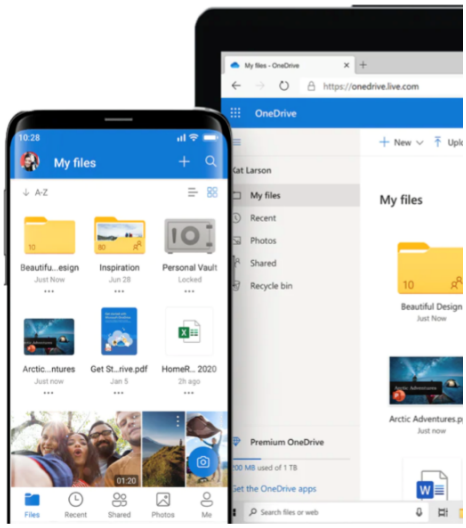
- A web-application (web interface)
- Via shared folders that hold a reference to the files stored in the cloud

Files can be synchronized

- a copy resides on a local computer hard drive is copied to the cloud
- A file residing in the cloud can be copied to a local computer hard drive
- Copying occurs automatically when files are changed
  - difference in files could result in conflicts a user must choose to which file to keep

Files can be versioned (you can recover older versions of a files)

- Older files may retain for 30 days and be automatically deleted



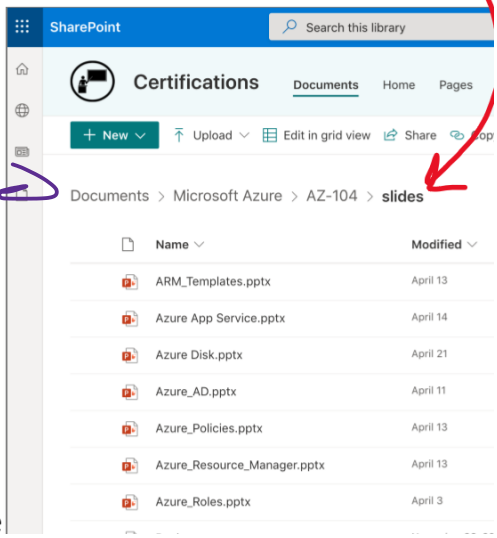
Microsoft 365 - SharePoint

Microsoft 365 SharePoint is **a web-based collaborative platform** that integra
Microsoft Office. Intended **for document management and shared storage**.

SharePoint Sites

Data within SharePoint organized around Sites
A site a collaborative space for teams with the following components:

- **Document Library**
- Pages
- Web Parts
- And more....

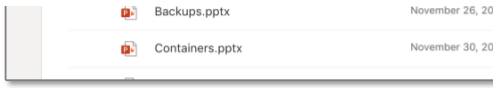


SharePoint Document Library

A document library is file storage and synchronization but ***designed for teams***. It is very similar to OneDrive, but files are owned by the company and not an individual.

You can apply robust permissions to access files within or outside

your organization



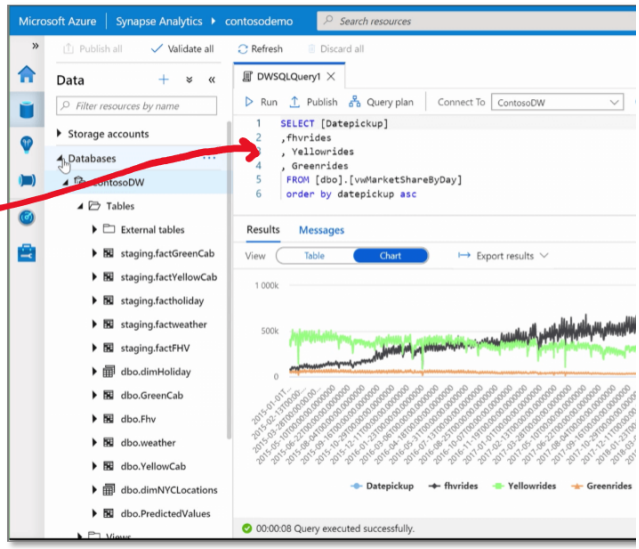
A Site always has default Document Library called “documents”

Azure Synapse Analytics

Azure Synapse Analytics is a **data warehouse** and **unified analytics platf**

Build ETL/ELT processes:
- in a code-free visual environment
- easily ingest data from more than 95 native connector

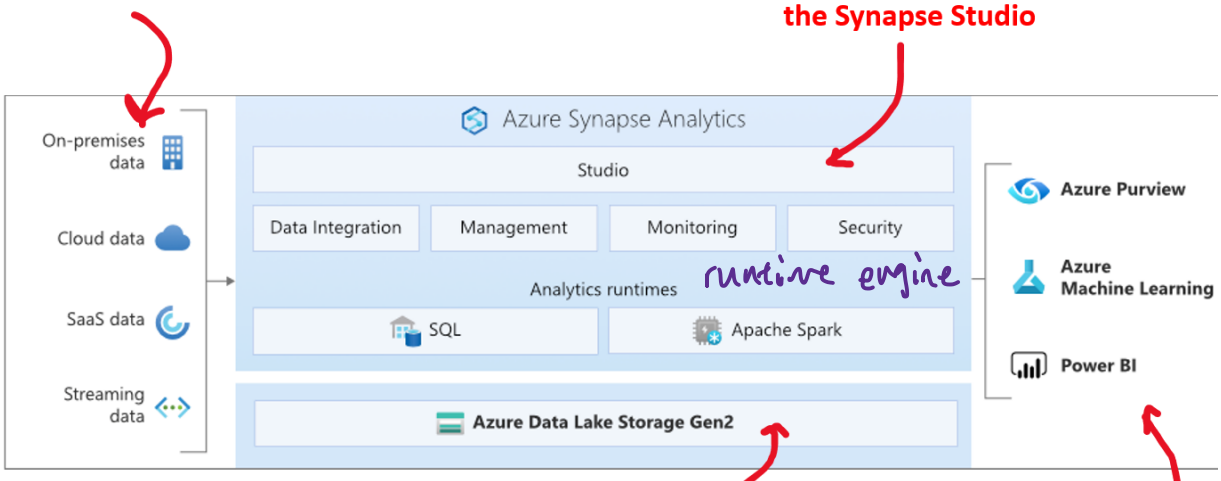
- Deeply integrated Apache Spark
- use T-SQL queries on both your data warehouse and Spark engines
- Supports multiple languages: T-SQL, Python, Scala, Spark SQL, and .Net
- Integrated with Artificial Intelligence (AI) and Business Intelligence tools (BI)
  - Azure Machine Learning
  - Azure Cognito Services
  - Microsoft Power BI



Azure Synapse Analytics

You can ingest data from many data sources.

Azure Synapse Analytics Is interfa the Synapse Studio



The data is stored in Object Storage Via Data Lake Storage Gen 2

You can outpu Azure Service

Synapse SQL

Synapse SQL is a distributed version of T-SQL designed for data warehouse wor

- extends T-SQL to address streaming and machine learning scenario
- use built-in **streaming** capabilities to land data from cloud data sources into SQL tables
- Integrate AI with SQL by using ML models to score data using the T-SQL PREDICT functio
- and offers both **serverless** and **dedicated** resource models

For **unpredictable** workloads (unplanned or bursty) use the **always-available, serverless SQL endpoint.**

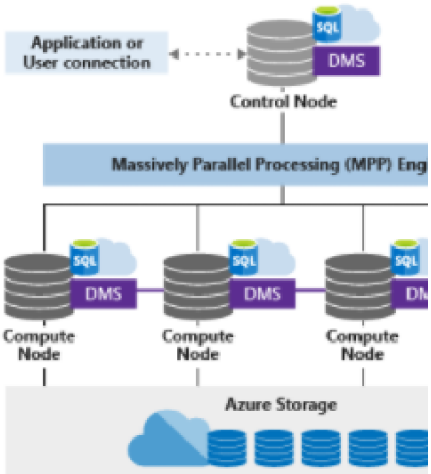
For **predictable** workloads
- create **dedicated SQL pools** to re processing power for data stored

Dedicated SQL Pool

Dedicated SQL pool is a query service over the data in your **data warehouse**

The unit of scale is an abstraction of compute power that is known as a **data warehouse unit (DWU).**

Once your dedicated SQL pool is created, you can import big data with simple PolyBase T-SQL queries, and then use the power of the distributed query engine to run high-performance analytics



Serverless SQL Pool

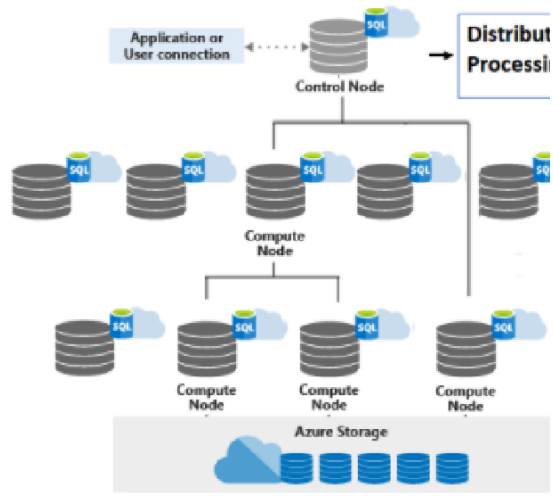
Serverless SQL Pool

Serverless SQL pool is a query service over the data in your **data lake**

Scaling done automatically to accommodate query resource requirements.

As topology changes over time by adding, removing nodes or failovers

it adapts to changes and makes sure your query has enough resources and finishes successfully.



Apache Spark for Synapse

Azure Synapse can **deeply and seamlessly integrates** with Apache Spark

- ML models with SparkML algorithms and AzureML integration for Apache Spa
  - with built-in support for Linux Foundation Delta Lake.
- Simplified resource model that frees you from having to worry about managir
- Fast Spark start-up and aggressive autoscaling.
- Built-in support for .NET for Spark allowing you to reuse your C# expertise anc code within a Spark application.

Apache Spark with Data Lake

Azure Synapse removes the traditional technology barriers between using SQL a
together. You can seamlessly mix and match based on your needs and expertise.

- Tables defined on files in the data lake are seamlessly consumed by either Spa
- SQL and Spark can directly explore and analyze Parquet, CSV, TSV, and JSON file
  data lake.
- Fast, scalable data loading between SQL and Spark databases

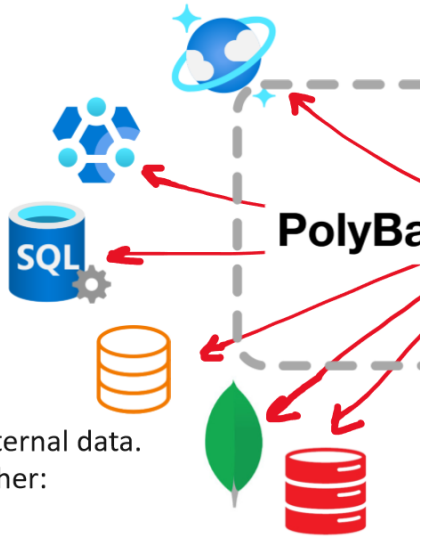
PolyBase

PolyBase is a **data virtualization feature** for **SQL Server**

PolyBase enables your SQL Server instance to query data
with T-SQL directly from:

- SQL Server
- Oracle
- Teradata
- MongoDB
- Hadoop clusters
- Cosmos DB

without separately installing client connection software.



PolyBase allows you to join data from a SQL Server instance with external data.
Prior to PolyBase to join data to external data sources you could either:

- Transfer half your data so that all the data was in one location.
- Query both sources of data, then write custom query logic to join and
  integrate the data at the client level.

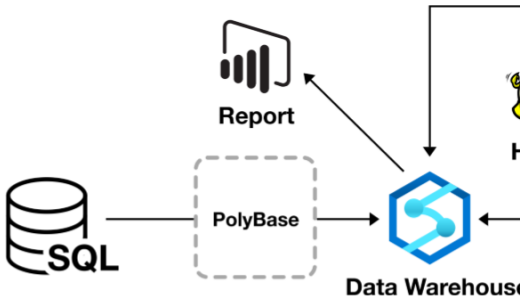
Azure Synapse Analytics - ELT

You can **perform ELT using Synapse SQL** in Azure Synapse Analytics.

The the fastest and most scalable way to load data is through

The the fastest and most scalable way to load data is through PolyBase external tables and the COPY statement

With PolyBase and the **COPY statement**, you can access external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.



The basic steps for implementing ELT are:

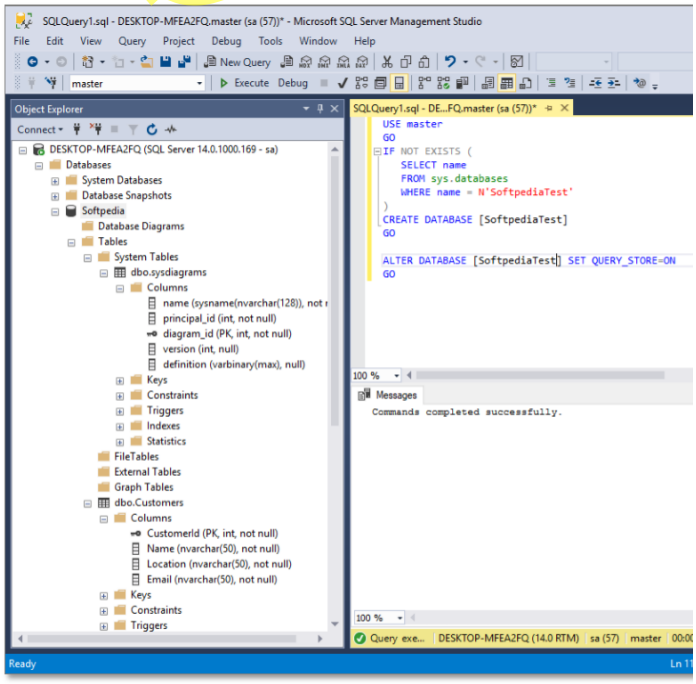
1. Extract the source data into text files.
2. Load the data into Azure Blob storage or Azure Data Lake Store.
3. Prepare the data for loading.
4. Load the data into staging tables with PolyBase or the COPY command.
5. Transform the data.
6. Insert the data into production tables.

SQL Server Management Studio (SSM

SQL Server Management Studio (SSMS) is an IDE for **managing any SQL infrast**

Access, configure, manage, administer, and develop all components of

- SQL Server
- Azure SQL Database
- Azure Synapse Analytics



Object Explorer
view and manage all of the objects in one or more instances of SQL Server

TemplateExplorer
build and manage files of boilerplate text that you use to speed the development of queries and scripts

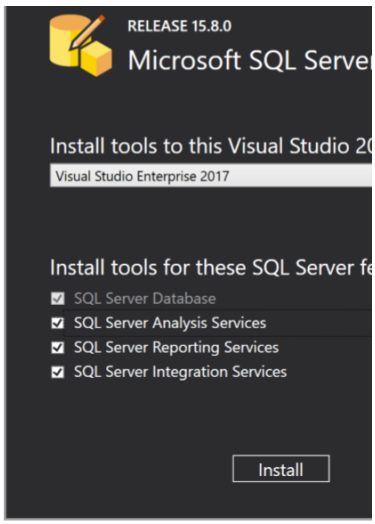
Solution Explore ***(deprecated)***
build projects used to manage administration items such as scripts and queries

SQL Server Data Tools (SSDT)

SQL Server Data Tools (SSDT) **transforms database development** by introducing a ubiquitous, dec that spans all the phases of database development inside Visual Studio

use SSDT Transact-SQL to build, debug, maintain, and

refactor databases.



SSDT also provides a visual Table Designer for creating and editing tables in either database projects or connected database instances
Be able to view control data related files
Easy to publish to SQL Database or SQL Server

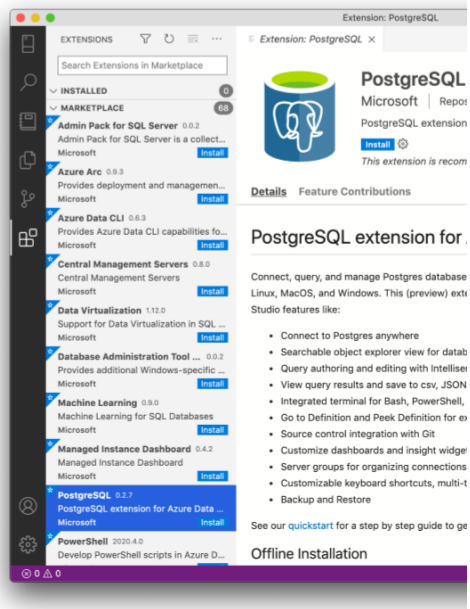
SQL Server Object Explorer in Visual Studio offers a view of your database objects similar to SQL Server Management Studio (SSMS)

- allows you to do light-duty database administration and design work
- easily create, edit, rename and delete tables, stored procedures, types, and function
- edit table data, compare schemas, or execute queries by using contextual menus right

Azure Data Studio

Azure Data Studio is a **cross-platform database tool** for data professionals using **on-p any cloud data platforms** for **Windows, macOS, and Linux**.

Query, design, and manage your databases and data warehouses



Azure Data Studio offers:

- a modern editor experience with IntelliSense
  - **Very similar experience to Visual Studio Code.**
- code snippets
- source control integration
- integrated terminal
- built-in charting of query result sets
- customizable dashboards
- Jupyter Notebooks connected to your datasets
- A marketplace of free extensions
  - SQL Database Inspector (inspect data with just a few clicks)
  - Kusto (KQL) extension for Azure Data Studio
  - PostgreSQL extension for Azure Data Studio
  - And many many more!

Business Intelligence (BI)

Business intelligence (BI) is both a data-analysis strategy and **technology** for business information.

The most popular BI tools are:

- Tableau
- Microsoft Power BI
- Amazon QuickSight

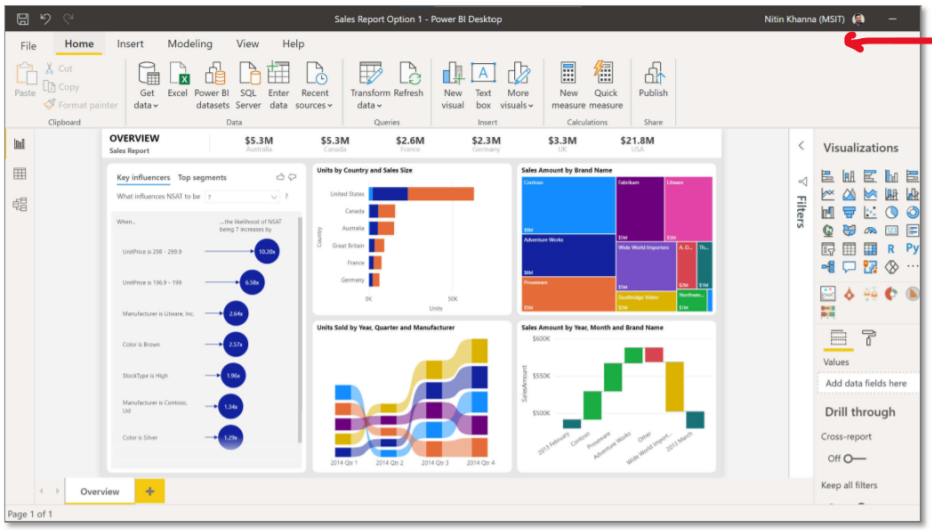
Business intelligence (BI) helps organizatio data-driven decisions by (BI) combining:

- business analytics
- data mining
- data visualization
- data tools
- infrastructure

- best practices

Microsoft Power BI

Power BI is a Business Intelligence tool for **visualization business data**.



Power BI Desktop
A way to design an

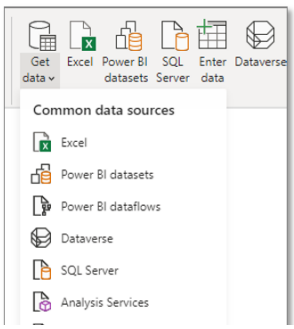
Power BI Mobile
View reports on th

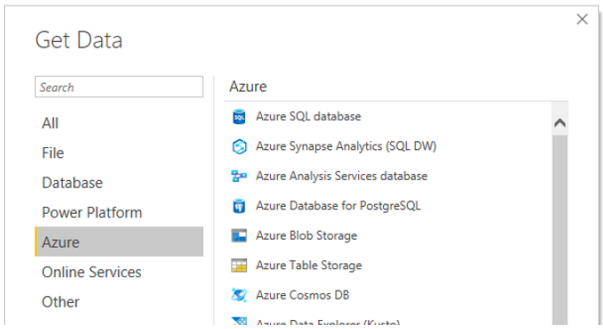
Power BI Service
Access and modify
cloud

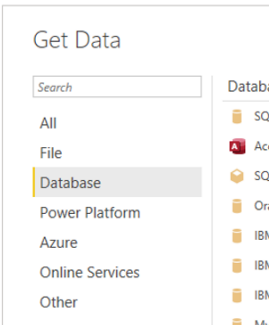
Power BI embedde
A way to embed Pc
components into y

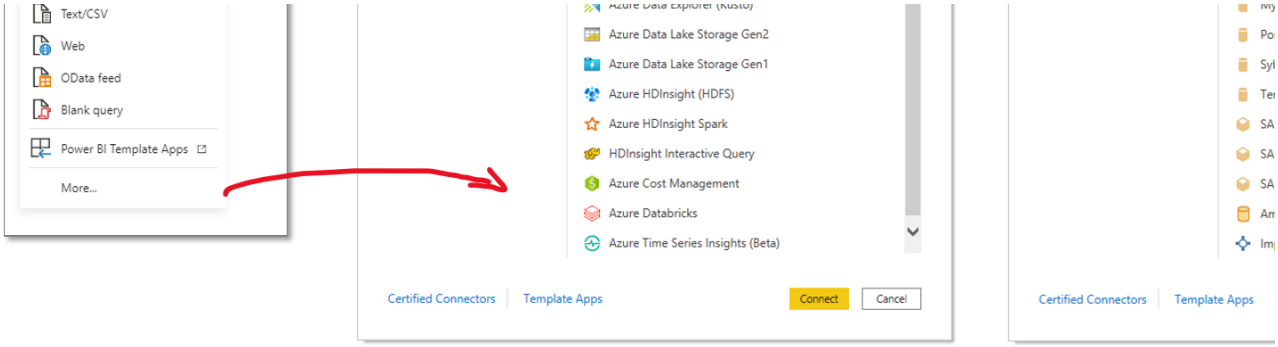
Microsoft Power BI Desktop – Data Sou

Microsoft Power BI can ingest data from **many data sources**.



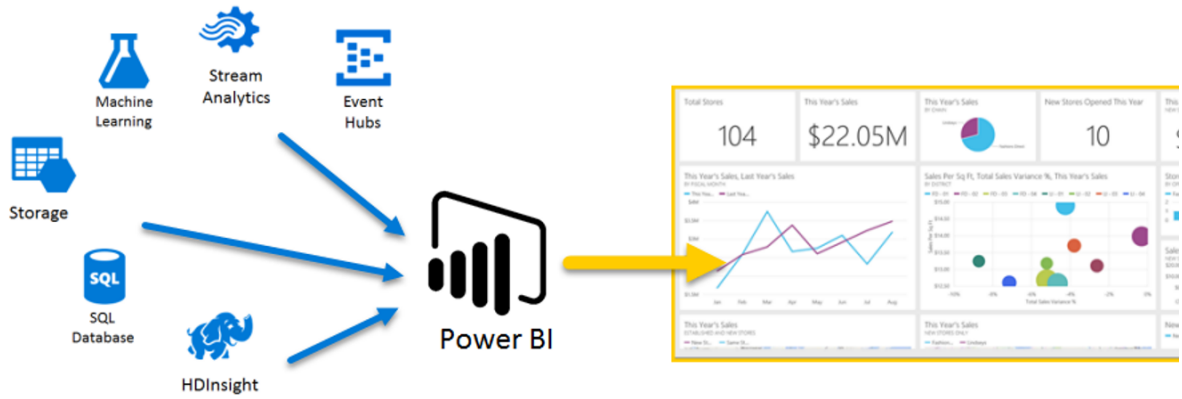






Azure and Power BI

Power BI can directly integrate with Azure Services:



Power BI Desktop vs Power BI Servic

Power BI Desktop
downloaded as a free *Windows* application and installed on a local *Windows* computer

Report designers use the Desktop application to publish the Power BI reports to the Power BI Service

Power BI Service
cloud-based service where use
interact with the repo

Users in the Power BI Service c
reports and create visuals based
data model and they can share a
with co-workers

Power BI Desktop	Both	Power BI Service
• Many data sources	• Reports	• Some data sources
• Transforming	• Visualizations	• **Dashboards**

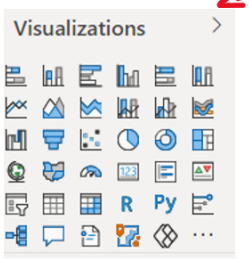
- Shaping and modeling
- Measures
- Calculated columns
- Python
- Themes
- RLS creation

- Security
- Filters
- Bookmarks
- Q&A
- R Visuals

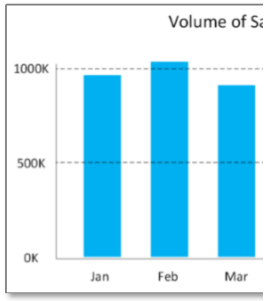
- Apps and workspace
- Sharing
- Dataflow creation
- Paginated reports
- RLS management
- Gateway connectior

Data Visualizations and Chart Types

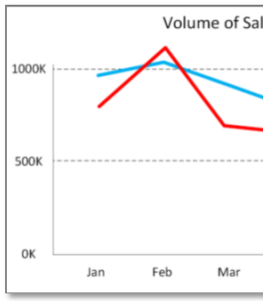
Power BI has **many kinds of visualizations**. We'll cover the most common ones.



Bar and column charts
See how a set of variables changes across different categories



Line charts
overall shape of an entire series of values



Data Visualizations and Chart Types

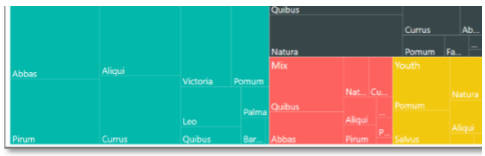
Matrix
tabular structure that summarizes data

Quarter Year	Q1 Revenue	YTD Revenue	Q2 Revenue	YTD Revenue
2015	$45,186	$45,186	$70,609	$115,795
2016	$52,154	$52154	$73,542	$125,696
2017	$51,388	$51,388	$68,149	$118,537
2018	$48,281	$48,281	$66,853	$115,134
2019	$53,145	$53,145	$49,135	$102,280

Key influencers
the major contributors to a selected result or value.

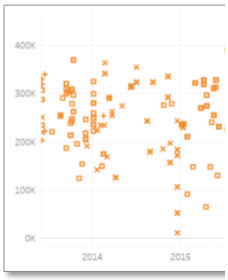
Treemap
charts of colored rectangles, with size representing the relative value of each item

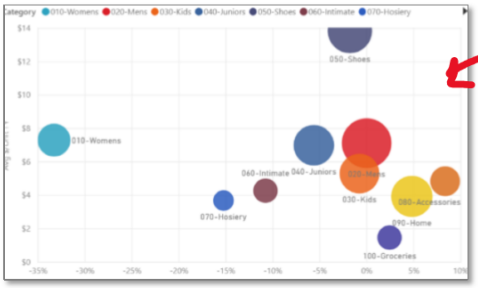
representing the relative value of each item



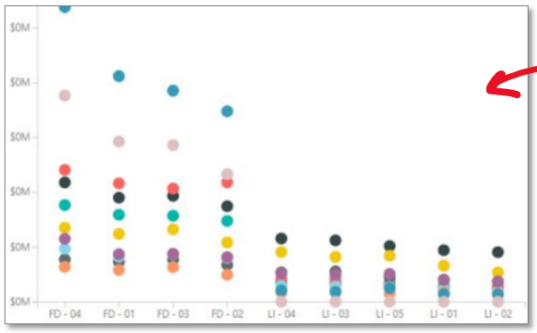
Data Visualizations and Chart Types

Scatter
relationship between two numerical values (X and Y Axis)
A bunch of dots on a graph



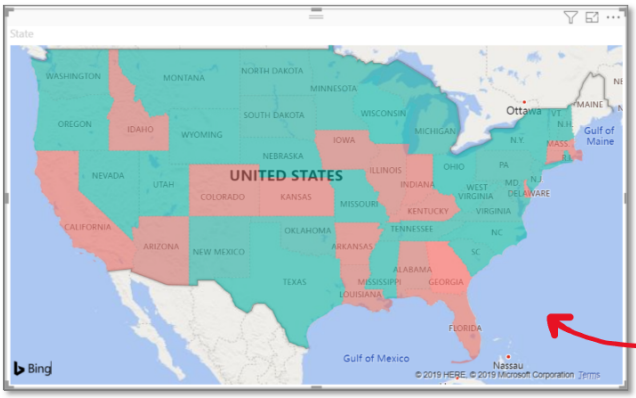


Bubble Chart
a scatter chart that replaces data points with bubbles, larger bubbles representing a third dimension



Dot plot Chart
similar to a bubble chart and scatter chart, but can plot categorical data along the X-Axis.

Data Visualizations and Chart Types



Filled map
A geographic map where different ar
can be filled in eg. States different co
or ranges of colors

Power BI Embedded

Azure Power BI Embedded is a platform-as-a-service (PaaS) analytics eml
that allows you to quickly embed **visuals**, **reports** and **dashboards** into ar

For independent software vendors (ISV)
enables you to visualize application data, rather than building that service yourself.

For developers
embed reports and dashbo
application for their custom

To use Azure Power BI embedded

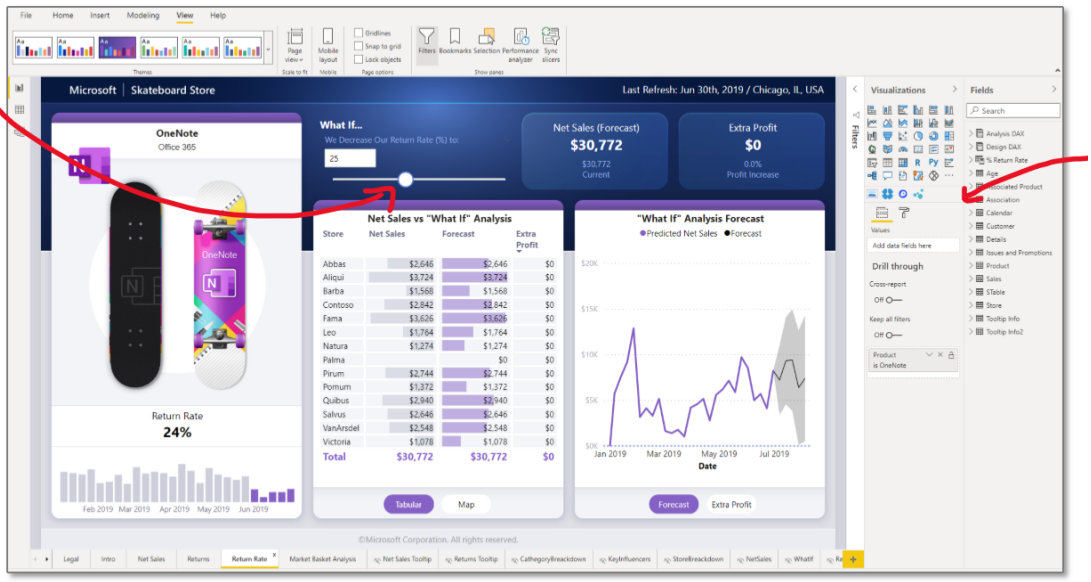
- You need a Power BI Pro user
- You need to create an App workspace
- You need to choose a capacity
  - Billing works via a capacity-based, hourly metered mode

Power BI – Interactive Reports

Power BI allows you to generate **reports which are interactive**

One way you can reports interactive is by **having knob and controls** directly in the report

Reports can be highly **stylized**



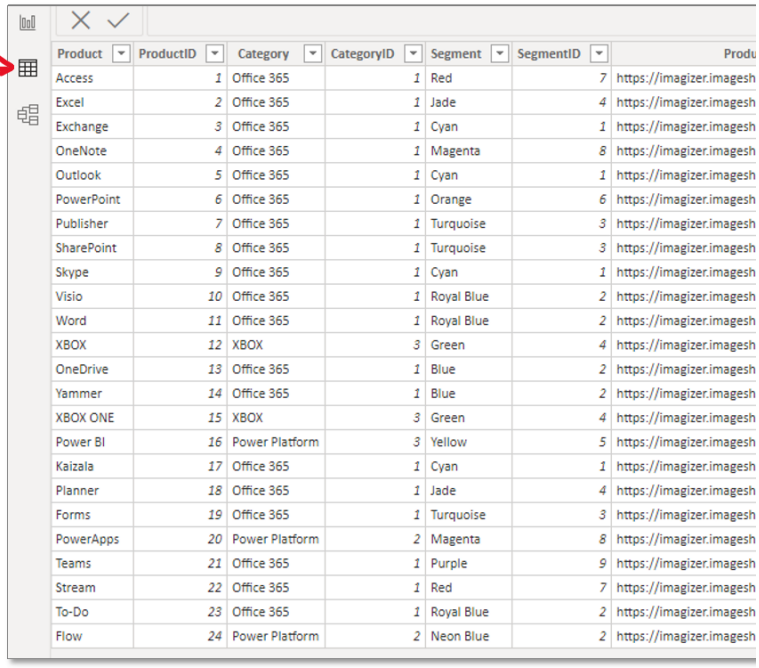
A report can contain may **pages**

Power BI – Interactive Reports

You can view the **underlying data** for an Interactive Report

(you can't do this with Dashboards)

PowerBI desktop



Product	ProductID	Category	CategoryID	Segment	SegmentID	Produ
Access	1	Office 365	1	Red	7	https://imagizer.imagesh
Excel	2	Office 365	1	Jade	4	https://imagizer.imagesh
Exchange	3	Office 365	1	Cyan	1	https://imagizer.imagesh
OneNote	4	Office 365	1	Magenta	8	https://imagizer.imagesh
Outlook	5	Office 365	1	Cyan	1	https://imagizer.imagesh
PowerPoint	6	Office 365	1	Orange	6	https://imagizer.imagesh
Publisher	7	Office 365	1	Turquoise	3	https://imagizer.imagesh
SharePoint	8	Office 365	1	Turquoise	3	https://imagizer.imagesh
Skype	9	Office 365	1	Cyan	1	https://imagizer.imagesh
Visio	10	Office 365	1	Royal Blue	2	https://imagizer.imagesh
Word	11	Office 365	1	Royal Blue	2	https://imagizer.imagesh
XBOX	12	XBOX	3	Green	4	https://imagizer.imagesh
OneDrive	13	Office 365	1	Blue	2	https://imagizer.imagesh
Yammer	14	Office 365	1	Blue	2	https://imagizer.imagesh
XBOX ONE	15	XBOX	3	Green	4	https://imagizer.imagesh
Power BI	16	Power Platform	3	Yellow	5	https://imagizer.imagesh
Kaizala	17	Office 365	1	Cyan	1	https://imagizer.imagesh
Planner	18	Office 365	1	Jade	4	https://imagizer.imagesh
Forms	19	Office 365	1	Turquoise	3	https://imagizer.imagesh
PowerApps	20	Power Platform	2	Magenta	8	https://imagizer.imagesh
Teams	21	Office 365	1	Purple	9	https://imagizer.imagesh
Stream	22	Office 365	1	Red	7	https://imagizer.imagesh
To-Do	23	Office 365	1	Royal Blue	2	https://imagizer.imagesh
Flow	24	Power Platform	2	Neon Blue	2	https://imagizer.imagesh

Power BI – Interactive Reports

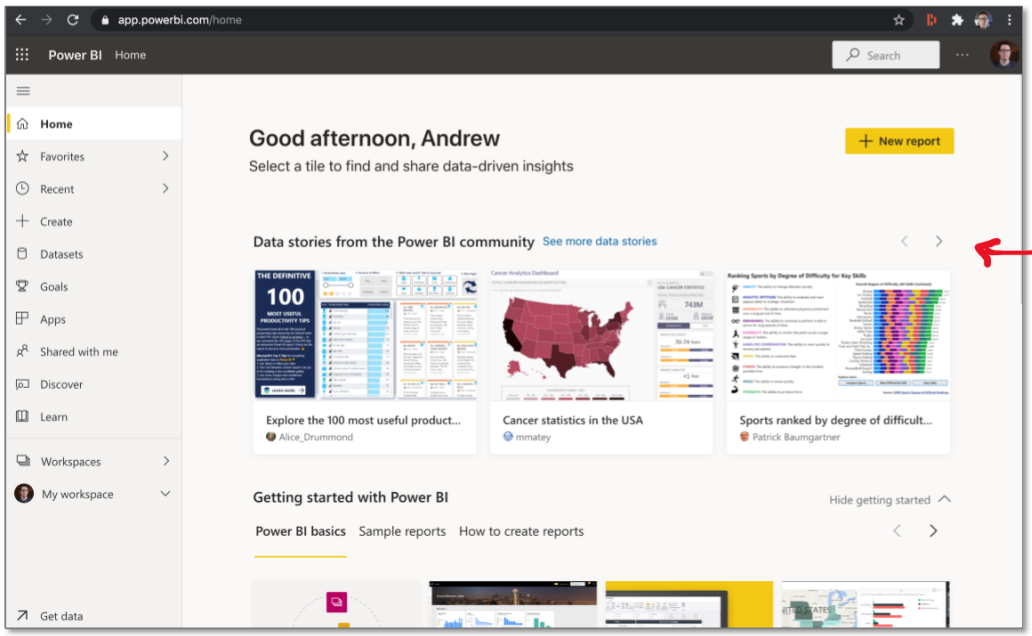
You can view **and modify the underlying data models** for a report

(you can't do this with Dashboards)

PowerBI desktop

Power BI Service

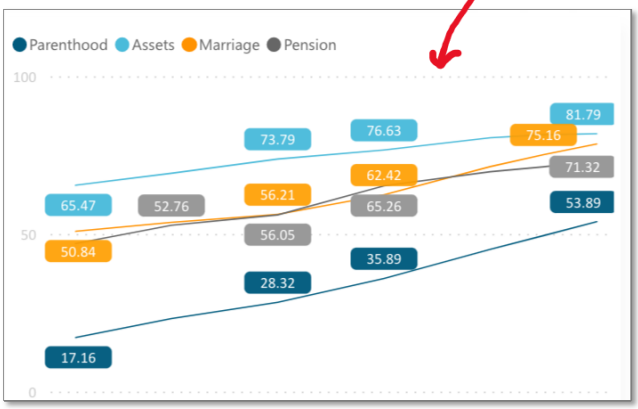
Power BI is **cloud-based service** where users view and interact with the rep
and where they can create **Dashboards**



You access Pov
by visiting app

Power BI – Dashboard Tiles

A **tile** is a **snapshot of your data**, pinned to the dashboard.

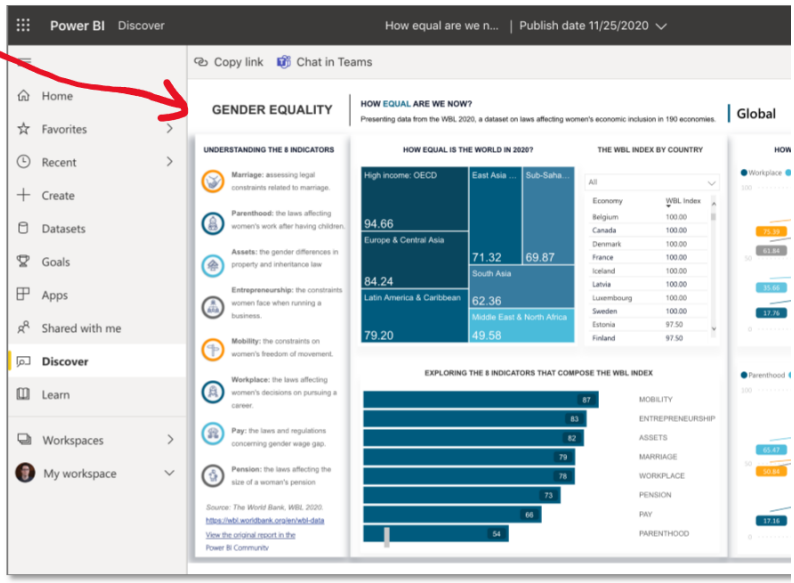


A tile can be created from a:

- **Report**
- **Dataset**
- Dashboard
- Q&A box
- Excel
- SQL Server Reporting Services (SS
- and more

Power BI Service – Dashboard

Power BI dashboard is a **single page**, often called a canvas, that tells a story through visualizations



The visualizations you see on the dashboard are called **tiles**.

You **pin** tiles to a dashboard from reports

Power BI – Reports vs Dashboards

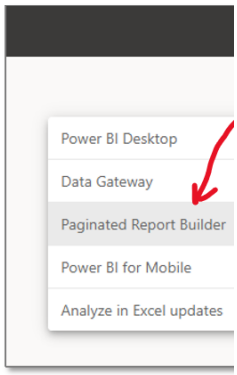
Capability	Dashboards	Reports
Pages	One page	One or more pages
Data sources	One or more reports and one or more datasets per dashboard	A single dataset per repo
Filtering	Can't filter or slice	Many different ways to f slice
Set alerts	Can create alerts to email you when the dashboard meets certain conditions	No
Feature	Can set one dashboard as your featured dashboard	Can't create a featured r
Can see underlying dataset tables and fields	No. Can export data but can't see the dataset tables and fields in the dashboard itself	Yes. Can see dataset tabl values that you have per
Customization	No	Can filter, export, view re add bookmarks, generat analyze in Excel, and mo

Paginated Reporting (RDL)

Paginated Reports a reports designed to fit into page format so they can be **printed** or **shared**
The data display of all data are tables which can span multiple pages.

Report Definition Language (RDL)

An XML representation of a SQL Server Reporting Services report definition
.A report definition contains data retrieval and layout information for a report
Paginated Reports are just a visulzation of an .rdl file.

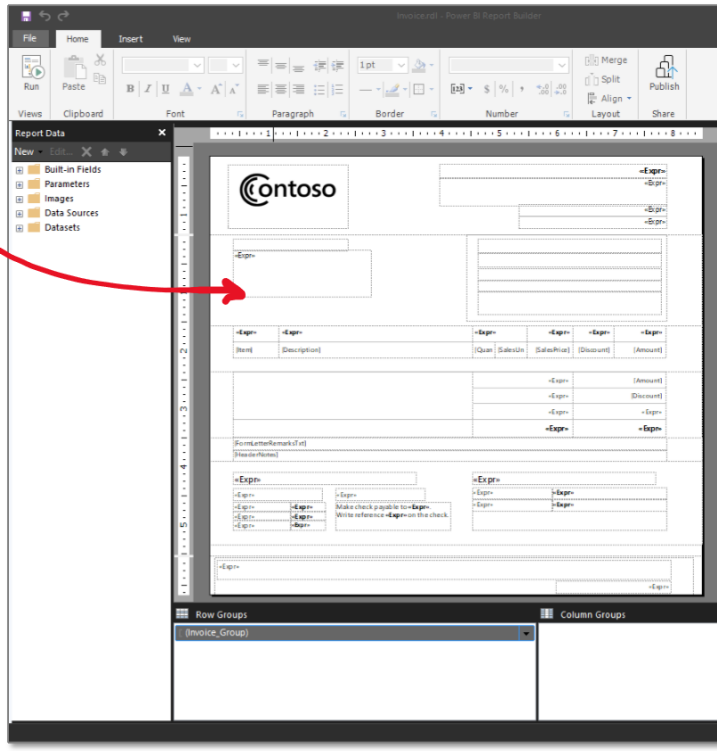


Power BI Report Builder
To design pixel-perfect paginated reports, you use Power BI Report Builder.
It is a tool specifically for the creation of Paginated Reports

You can downlo
from Power BI S

Paginated Reporting (RDL)

This is what a **Paginated Report**
Looks like in Power BI Report Builder

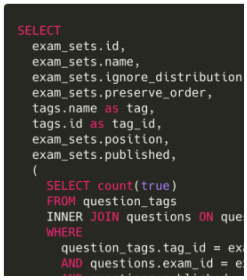


Structured Query Language

Structured Query Language (SQL) designed to **access and maintain data for a relational database management system (RDBMS)**

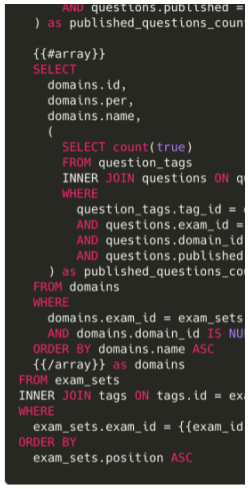
We use SQL to get to **insert, update, delete, view** data from our database's tables.

SQL can join many tables and include many functions



```
SELECT
  exam_sets.id,
  exam_sets.name,
  exam_sets.ignore_distribution
  exam_sets.preserve_order,
  tags.name as tag,
  tags.id as tag_id,
  exam_sets.position,
  exam_sets.published,
  (
    SELECT count(true)
    FROM question_tags
    INNER JOIN questions ON que
    WHERE
      question_tags.tag_id = ex
      AND questions.exam_id = e
```

SQL can join many tables and include many functions to transform the final outputted result



The SQL Syntax was standardized (ISO 9075).

Relational databases will mostly adhere to this standard while adding in their own database specific features.

SQL is a highly transferable skill and we see SQL being used in Non-Relational databases to provide a popular and familiar querying tool.

SQL Syntax

Clauses

- WHERE Clause
- ORDER BY clause
- HAVING Clause
- UPDATE Clause
- GROUP BY Clause

< MAYEB JUST SKIPT THIS S

Expressions
Scalar or tabular data that is used along side a Predicate to provide conditional data to be returned

Predicates
The condition of how to filter the data

Queries

Statements

OLAP vs OLTP

Online **Transaction** Processing (OLTP) VS. Online **Analytical** Processi

Database
A database is built to store current transactions and enable **fast access to specific transactions for ongoing business processes** e.g SQL servers

Data Warehouse
A data warehouse is built to store l of historical data and **enable fast, c queries across all the data**

small transactions

- Single Data Source
- **Short transactions** (small and simple queries)
  - with an emphasis on writes.
- Many transactions
- Latency sensitive
- Small payloads

Use Case: (General Purpose)
Adding Items to your shopping cart

- Multiple Data Sources
- **Long transactions** (long and c
  - with an emphasis on read
- Few transactions
- Throughput sensitive
- Large payloads

Use case: (Analytics)
Generating Reports

Open-Source Relational Databases

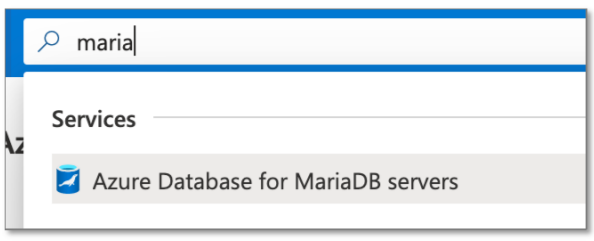
Created by **MySQL AB** then acquired by **Sun Microsystems** and then acq
MySQL is a **pure relational database (RDBMS)**
It is a simpler database which makes it **easy to setup, use and maintain**
Has multiple storage engines: InnoDB and MyIsam
The **most popular** relational database

MariaDB is a fork of MySQL by the original creators of **MySQL AB**
After Oracle acquired MySQL there was concern that Oracle may change open-source licensing or stop future MySQL being free to use.

PostgreSQL evolved from the Ingres project at the University of Californi
Postgres is an **object-relational database (ORDBMS)**
Just has a single storage engine Ingres engine
The **most advanced** relational database
- Full text search, table inheritance, triggers, rows and data types requi

Azure Database for Maria DB, PostgreSQL an

If you want to **deploy** an open-source database on Azure you need to search by its nam

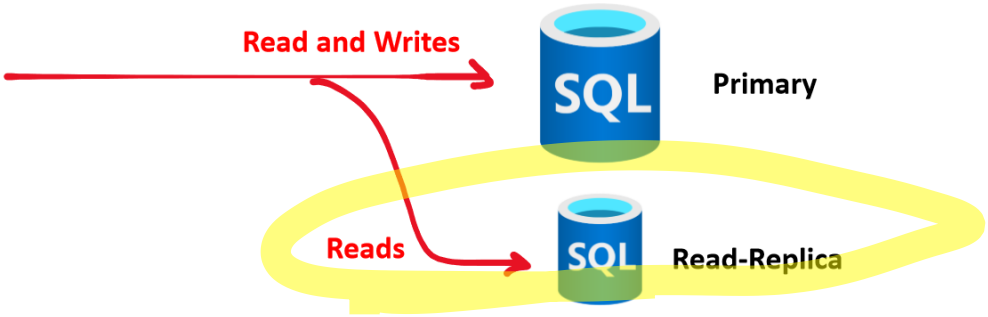


Azure DB Read Replicas

What is a read replica?
A read replica is a copy of your database that is kept synced your primary database.
This additional database is used to improve **read contention** by offloading reads to a databas
dedicated to perform read operations

Read Replicas are available for:
- Azure SQL Database
- Azure SQL Managed Instance

You can have multiple read-replicas for a dat



In simple use cases a Read Replica can act as an OLAP for a relational databases that is very si

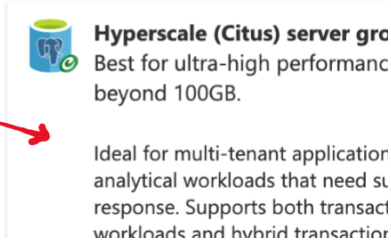
Citus on Azure

Citus is an **open-source Postgres extension** that transforms Postgres into a **distributed database**.

Distributed dat

Citus extends postgres to provide better support for:
- Database sharding (easy horizontal saling)
- Realtime queries (great for real-time analytics dashboards)
- Multi-tenany (great for SaaS company's)
- Time series workloads

Azure Postgres for **HyperScale deploy option** is just a managed postgres cluster us the Citus extension

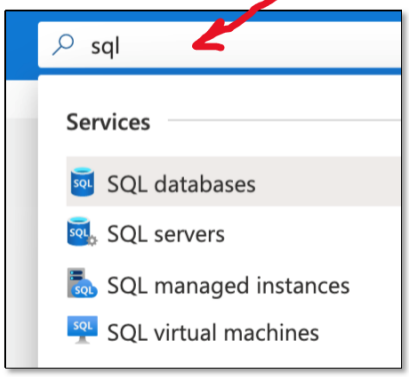


workloads.

Create Learn more

Azure SQL Family

Azure has **multiple solutions** for relational databases



SQL Server on Azure Virtual Machines
- When you need OS-level control and access
- When you need to lift-and-shift your workloa
- When you have existing SQL licenses and you Azure Hybrid Benefit

SQL Managed Instance
- When you have an existing database that you
- broadest SQL server engine compatibility
- highly available, disaster recovery, automate
- Ideal for most migrations to the cloud

Azure SQL Database
- Fully managed SQL databases
- designed to be fault-tolerant
- built-in disaster recovery
- highly available
- designed to scale

SQL Servers
The underlying servers for Azure SQL D

Azure Elastic Pools

Azure SQL Database elastic pools are a simple, cost-effective solution for **managing and sc multiple databases** that have **varying and unpredictable usage demands**

Databases in an elastic pool are on:
- a single server
- share a set number of resources at a set price

Elastic pools in Azure SQL Database enable SaaS developers to optimize the price performance for a group of databases within a prescribed budget while delivering performance elasticity for each database.

Connectivity Architecture – SQL Database and Azure Syna

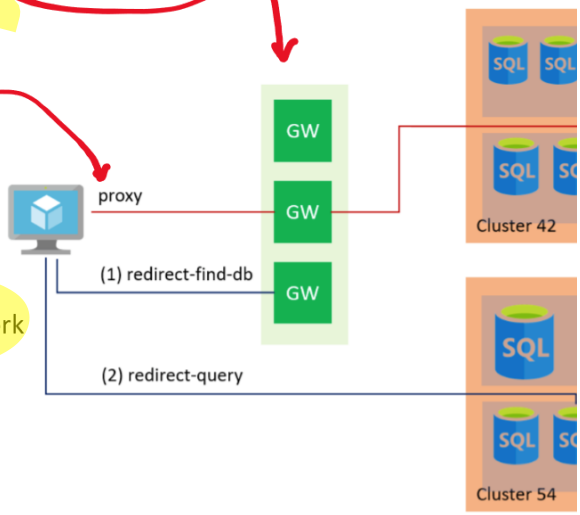
When a connection from a server to an Azure SQL database the client will connect to a **gateway** that listens on **port 1443**

Based on the **connection policy** the gateway will grant traffic and route access to the appropriate database

There are three kinds of connection policies:

Proxy
connections are proxied through a gateway
increased latency and reduced throughput
intended for workloads connecting from outside the Azure Network

Redirect (Recommended)
establishes a direct connection
reduced latency and improved throughput
intended for workloads connecting inside the Azure Network



Default
When you create a database it will default to either Proxy or Redirect
Based on workloads inside or outside the Azure networking

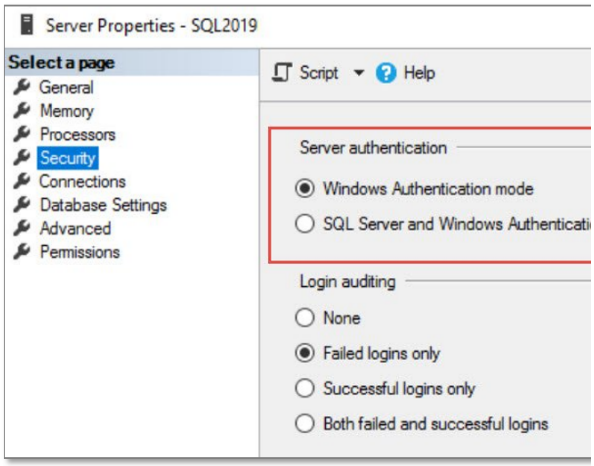
MS SQL Database Authentication

During setup of your MS SQL databases must **select an authentication mode**

- **Windows Authentication mode**
  - enables Windows Authentication and disables SQL Server Authenticatio
- **Mixed mode**
  - enables both Windows Authentication and SQL Server Authentication

Windows Authentication (recommended)
- Specific Windows user and group accounts are trusted to log in to SQL Server
- Very secure, and very easy to modify or revoke access

SQL Server Authentication
- A username and password is set and stored on the primary database
- cannot use Kerberos security protocol
- login password must be passed over the network at the time of the connection (additional attack points)
- easier to connect to database from outside a domain or from a web-based interface

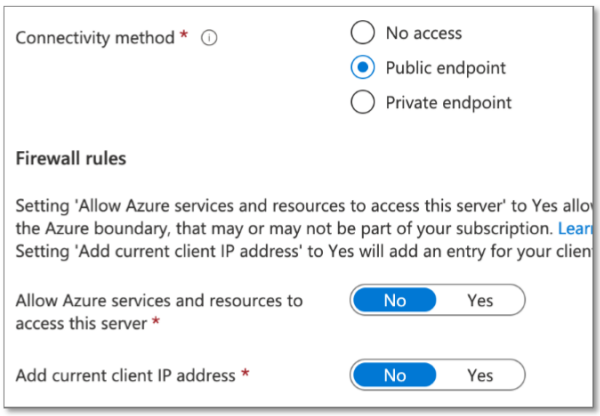


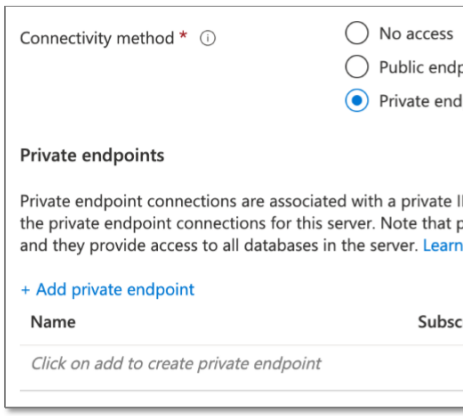
Network Connectivity

For SQL Database you can choose network connectivity to have a **Public** or **Private** Endpoin

Public Endpoints are reachable outside the Azure Network over the internet
- You would use **Firewall rules** to protect the database

Private Endpoints are only reachable within the Azure Network (or a connection originating from
- You would use **Azure Private Link** to keep traffic within the Azure Network.





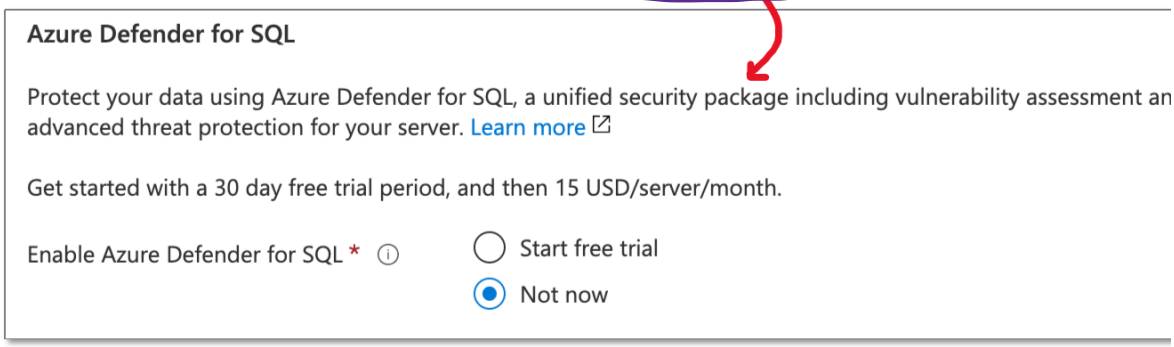
Azure Defender for SQL

Azure Defender for SQL is **a unified package for advanced SQL security c**
for **Vulnerability Assessment** and **Advanced Threat Protection**

Azure Defender is available for:
- Azure SQL Database
- Azure SQL Managed Instance
- Azure Synapse Analytics

What it does:
- discovering and classifying sensitive data
- surfacing and mitigating potential database v
- detecting anomalous activities

You can turn it on at anytime and you pay a monthly cost

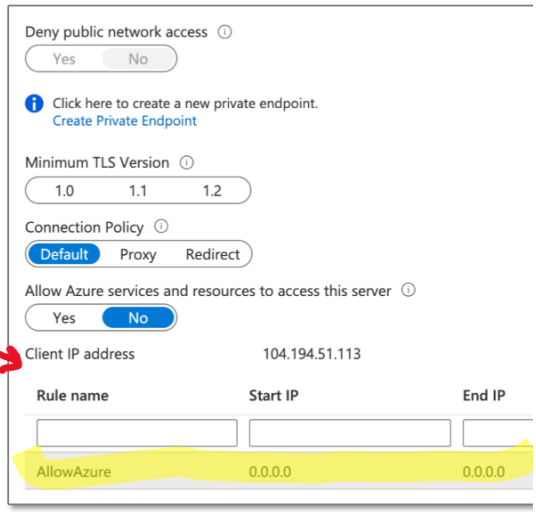


Azure Database Firewalls Rules

Diff from Azure firewall

Azure databases are protected by **server firewalls**
A server firewall is an internal firewall that resides on the database server
All connections are **rejected by default** to database

You can set server firewall rules within the Azure Portal



You can set server firewall rules via T-SQL

```
EXECUTE sp_set_database_firewall_rule N'OnlyAllowServer','0.0.
```

Always Encrypted

Always Encrypted is a feature that encrypts columns in an Azure SQL Database

If you had a column for credit cards you would want to Always Encry

Always Encrypted uses two types of keys:
- column encryption keys — used to encrypt data in an encrypted column
- column master keys — a key-protecting key that encrypts one or more column encrypt

You apply Always Encrypted using T-SQL

Role-Based-Access-Controls (RBA

Role-Based-Access-Controls (RBAC) is when you can apply roles to users to
the fine-grade actions for specific Azure services

SQL DB Contributor
- manage SQL databases, but not access to them
- can't manage their security-related policies or their parent SQL servers

SQL Managed Instance Contributor
- manage SQL Managed Instances and required network configuration
- can't give access to others

SQL Security Manager
- manage the security-related policies of SQL servers and databases
- but not access to SQL servers

SQL Server Contributor
- manage SQL servers and databases
- but not access to them SQL servers

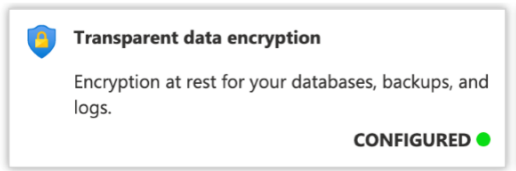
Transparent Data Encryption (TDE)

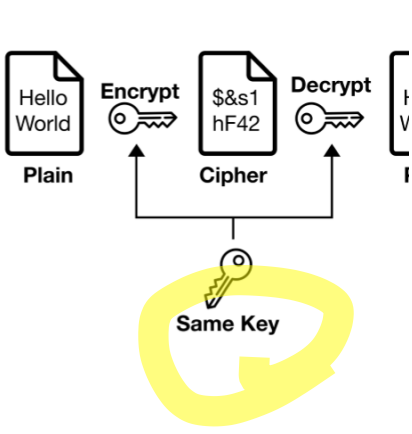
Transparent Data Encryption (TDE) **encrypts data-at-rest** for Microsoft Data

TDE can be applied to:
- SQL Server
- Azure SQL Database
- Azure Synapse Analytics

- TDE does real-time I/O encryption and decryption of d
- encryption uses a database encryption key (DEK)
- database boot record stores the key for availability du
- The DEK is a **symmetric key** (same cryptographic **keys** f the **encryption** of plaintext and the decryption of ciphe

Steps to apply TDE to a database:
- Create Database Master Key
- Create a Certificate to support TDE
- Create Database Encryption Key
- **Enable TDE on Database**



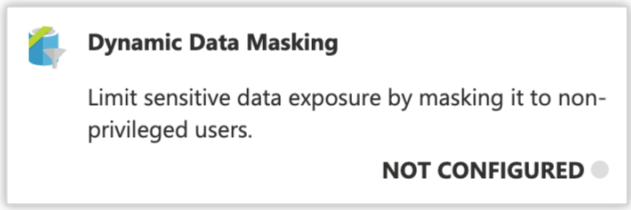


Dynamic Data Masking

Data Masking is when a **request for data is transformed to mask sensiti**

The data in the database remains untransformed

Dynamic Data Masking can be applied to:
- Azure SQL Database
- Azure SQL Managed Instance
- Azure Synapse Analytic



You create a Dynamic Data **Masking Policy**:
- SQL users excluded from masking — users who can get data unmasked
- Masking rules — what fields should be masked
- Masking functions — how to apply masking to fields

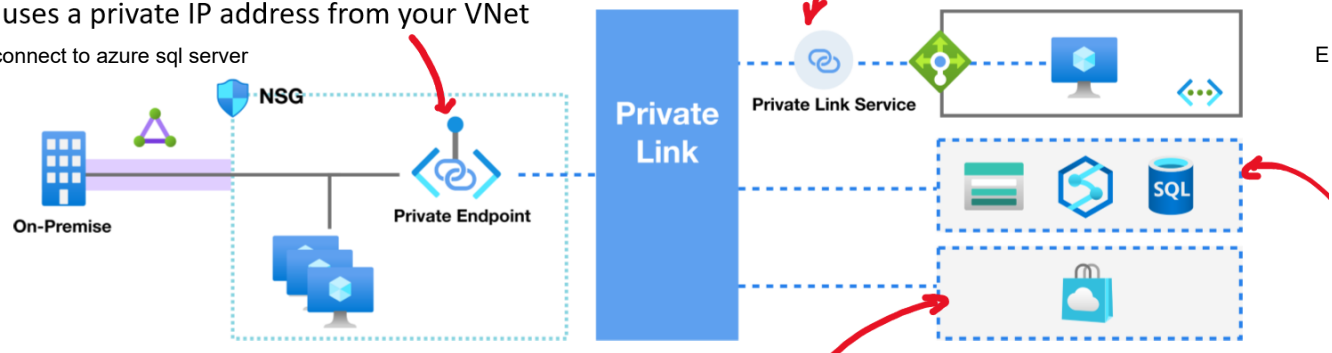
Private Links

Azure Private Links allows you to **establish secure connections** between Azu resources so traffic **remains within the Azure Network**

Private Link Endpoint is an **Network Interface** that connects you privately and securely to a service powered by Azure Private Link. Private Endpoint uses a private IP address from your VNet

Private Link Service allows you to c workload to Private Link. You need **Internal Load Balancer** and associat

E.g. connect to azure sql server



E

Third-Party provides can be powered by Private Link

Many Azure services by d Private Link eg. Azure Sto

T-SQL

Transact-SQL (T-SQL) is a set of programming extensions from Sybase and Mi that add several features to the Structured Query Language (SQL),

T-SQL expands on the SQL standard to include:
- procedural programming
- local variables

For Microsoft SQL Server there are five groups
- Data **Definition** Language (DDL)
  - used to define the database schema

- various support functions for string processing
- date processing
- Mathematics
- changes to the DELETE and UPDATE statements

- Data **Query** Language (DQL)
  - used for performing queries on the da
- Data **Manipulation** Language (DML)
  - manipulation of data in the database
- Data **Control** Language (DCL)
  - rights, permissions and other controls
- **Transaction Control** Language (TCL)
  - transactions within the database

Data Definition Language (DDL)

A Data **Definition** Language (DDL) is SQL syntax commands for
creating and modifying the database or database objects
eg: table, index, views, store procedure, function and triggers

CREATE
Create database or database objects
to create database and its objects

```
CREATE TABLE
    Id int,
    FullName
);
```

```
ALTER TABLE Users
ADD Email varchar(255);
```

ALTER
Alters the structure of an existing database

DROP
Delete objects from the database

```
DROP TABLE
```

```
TRUNCATE TABLE Users;
```

TRUNCATE
Remove all records from a table

COMMENT
Add comments

```
-- This table is for pe
SELECT * FROM Users;
```

execute

```
EXEC sp_rename 'Users', 'People';
```

RENAME
rename a database object.

only find in T-SQL

Data Manipulation Language (DML)

```
INSERT INTO Users VALUES
(1, 'Andrew', 'Brown',
'andrew@exampro.co');
```

INSERT
insert data into a table

UPDATE
updates existing data within a table

```
UPDATE Users
SET email = 'hello@
WHERE ID = 8;
```

```
DELETE FROM Users
WHERE Id = 6;
```

DELETE
delete all records from a table

MERGE – UPSERT T-SQL specific
To insert and update records at the same time

CALL
call a PL/SQL or Java subprogram.

```
CALL CalcDistnace 'To
```

LOCK TABLE
concurrency control to ensure two people are not writing to the program at the same time

Data Query Language (DQL)

```
SELECT
  Users.Id,
  Users.FullName,
  Users.Email
FROM Users;
```

SELECT
Select data from a table, And be able to specific exactly which columns to return

SHOW
Describe what a table looks like (what column it contains)

```
EXEC sp_columns U
```

EXPLAIN PLAN
Returns the query plan for a Microsoft Azure Synapse Analytics SQL statement without running the statement

```
EXEC sp_help Users
```

HELP
Reports information about a database object

Data Control Language (DCL)

```
GRANT SELECT, INSERT, UPDAT
DELETE ON employees TO andr
```

GRANT
allow users access privileges to database.

REVOKE
withdraw users access privileges given by using the GRANT command.

```
REVOKE DELETE ON employees FROM
bayko;
```

Transaction Control Language (TCL

Transaction Control Language commands are used to manage transactions in the database.

↓ need mu
things

COMMIT
set to permanently save any transaction into the database

ROLLBACK
restores the database to last committed state

SAVEPOINT
used to temporarily save a transaction so that you can rollback to that point whenever ne

SET TRANSACTION
specify characteristics for the transaction.

MS SQL Commands

SQL Commands				
DDL	**DML**	**DQL**	**DCL**	
CREATE	INSERT	SELECT	GRANT	
ALTER	UPDATE	SHOW	REVOKE	
DROP	DELETE	EXPLAIN PLAN		
TRUNCATE	MERGE	HELP		
COMMENT	UPSERT			
RENAME	CALL			
	LOCK TABLE			

Defining	Manipulating	Querying	Controlling

These are the two the SQL documents the exam will focus upon

What is a Key / Value store?

A key/value stores a **unique key** alongside a value

Key	Value
Data	10101010001010110010100101010001
Worf	01101011000101010101011100010
Ro Laren	00101010010101100101010101010

Key values stor
They generally
- Relationshi
- Indexes
- Aggeration
- transacti

Key	Value
Data	{species: android, rank: 'lt commander' }
Worf	{species: klingon, rank: 'lt commander' }
Ro Laren	{species: bajoran, affiliation: 'maquis'}

A simple key
interpret thi
dictionary (a
or hash)

A key/value store can resemble tabular data, it does not have to have the consistent columns per row (hence its schemaless)

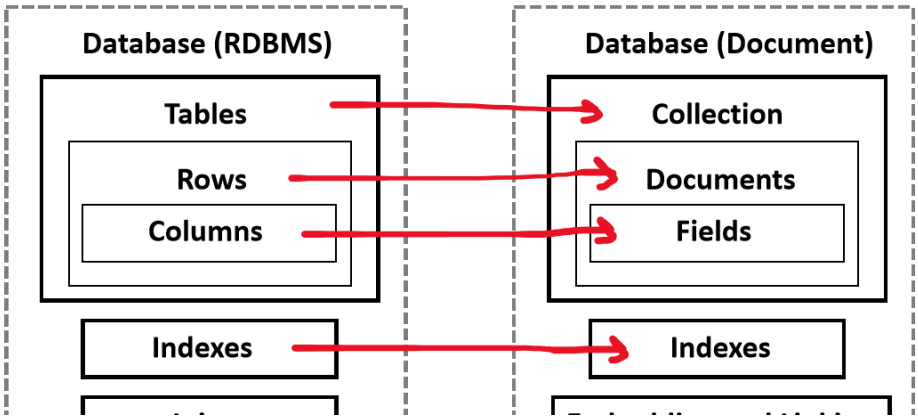
Key (Name)	Species	Rank	Affiliation
Data	andriod	Lt commander	
Worf	klingon	Lt commander	
Ro Laren	bajoran		maquis

Due to the
they can s
relational

What is a Document store?

A document store is a NOSQL database that stores **documents** as its primary data structure.
A document could be an XML but more commonly is JSON or JSON-Like
Document stores are sub-class of Key/Value stores

The components of a document store compared to Relational database



Joins → Embedding and Linking

MongoDB

BSON

MongoDB is an open-source document database which stores JSON-li
The primary data structure for MongoDB is called **BSON**

Binary JSON (BSON)

- BSON is a subset of JSON and so its data structure is very similar.
- BSON is designed to be **efficient both in storage space and scan-speed** compared to JSON
- BSON has more data-types than JSON:
  - Eg. Datetime, byte arrays, regular expressions, MD5 binary data, javascript code

```
BSON:
  \x16\x00\x00\x00          // total document si
  \x02                      // 0x02 = type String
  hello\x00                 // field name
  \x06\x00\x00\x00world\x00 // field value (size
  \x00                      // 0x00 = type EOO (
```

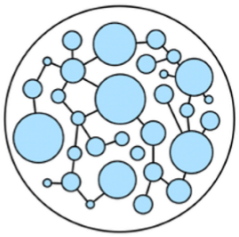
What it looks like to perform an operation on a MongoDB databases

```
db.inventory.insertMany([
   { item: "journal", qty: 25, size: { h: 14, w: 21, uom: "cm
   { item: "notebook", qty: 50, size: { h: 8.5, w: 11, uom: "
   { item: "paper", qty: 100, size: { h: 8.5, w: 11, uom: "in
   { item: "planner", qty: 75, size: { h: 22.85, w: 30, uom:
   { item: "postcard", qty: 45, size: { h: 10, w: 15.25, uom:
]);
```

MongoDB

- MongoDB supports searches against:
  - fields
  - ranged queries
  - regular-expressions
- MongoDB supports **primary** and **secondary** indexes
- High availability can be obtained via replica sets (replica to offload reads or acts a stand-by in c
- MongoDB scales horizontally using sharding
- MongoDB can run over multiple servers via load balancing
- MongoDB can be used as a file system, called **GridFS**
  - with load balancing and data replication features over multiple machines for storing files.
- MongoDB provides three ways to perform aggregation (grouping data during a query)
  - aggregation pipeline
  - map-reduce
  - single-purpose aggregation
- MongoDB supports fixed-size collections called capped collections
- MongoDB claims to support multi-document ACID transactions

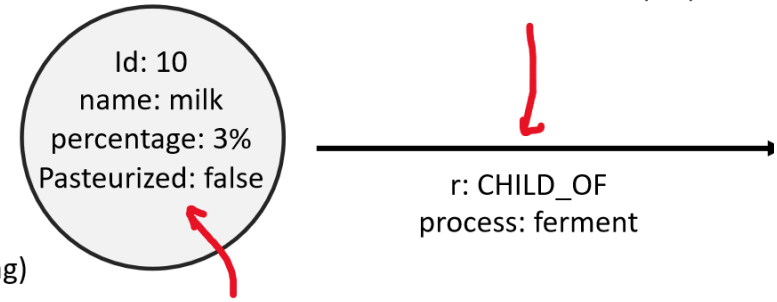
What is a Graph database?



A graph database is a database composed of a data structure that uses vertices (nodes, dots) which form **relationship** to other vertices via edges (arcs, lines)

Use Cases for Graph Database:

- Fraud detection
- Real-time recommendation engines.
- Master data management (MDM)
- Network and IT operations.
- Identity and access management (IAM)
- Traceability in Manufacturing
- Contact Tracing
- Data Lineage for GDPR
- Customer 360-degree analysis (marketing)
- Product recommendations
- Social Media graphing
- Feature Engineering (ML)



What is a Apache TinkerPop?

Apache TinkerPop is a is a **graph computing framework** for databases (OLTP) and graph analytic systems (OLAP)

TinkerPop enables developers to use **an vendor-agnostic distrusted framework** to traverse (query) many different graph systems:

Alibaba Graph Database	**Hadoop (Spark)**	**OrientDB**
Amazon Neptune	HGraphDB	OverflowDB
ArangoDB	Huawei Graph Engine Service	Apache S2Graph
Bitsy	HugeGraph	Sqlg
Blazegraph	IBM Graph	Stardog
CosmosDB	JanusGraph	TinkerGraph
ChronoGraph	JanusGraph (Amazon)	**Titan**
DSEGraph	**Neo4j**	Titan (Amazon)
GRAKN.AI	neo4j-gremlin-bolt	Titan (Tupl)
		Unipop

TinkerPop includes a graph traversal language called **Gremlin**
Which is the single language that can be used for all these graph systems

What is Gremlin?

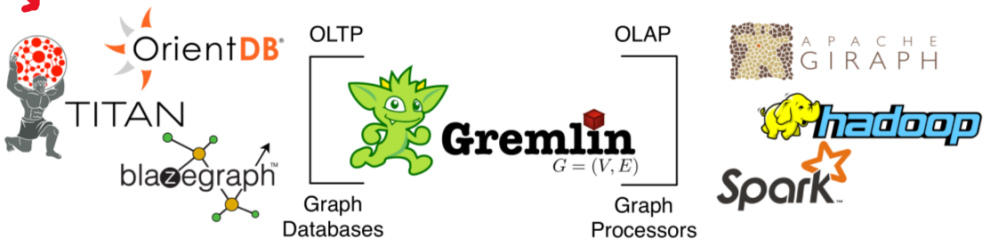
What is Gremlin?

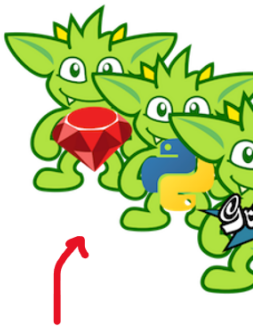
Gremlin is the **graph traversal language** for **Apache TinkerPop**

```
g.V().has("name","gremlin").as("a").
  out("created").in("created").
    where(neq("a")).
  groupCount().by("title")
```

Gremlin is design to the "Write once, run anywhere" (**WORA**)
Gremlin traversal can be evaluated as

- **real-time database query (OLTP)**
- or as a **batch analytics query. (OLAP)**





Gremlin Host Lar
means You can u
programming lan

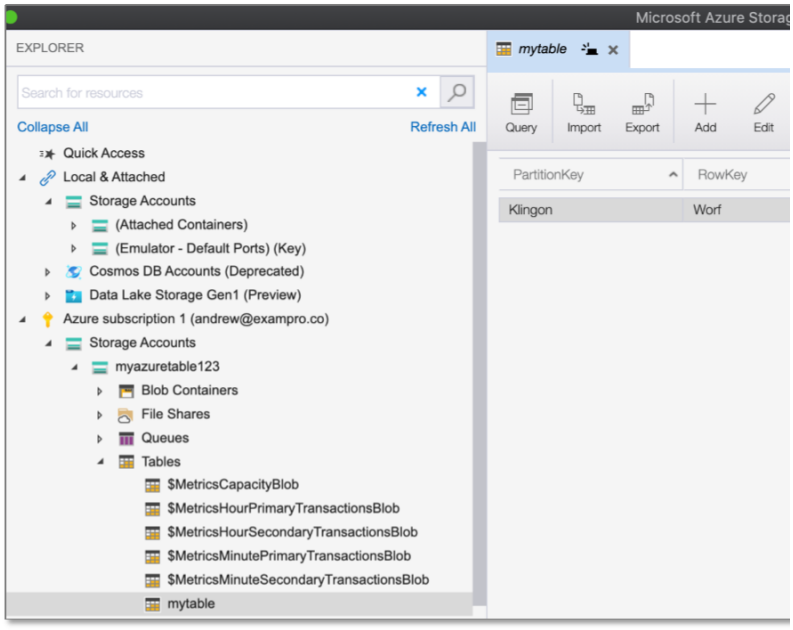
Azure Tables

Azure Table storage is a a **NoSQL key/value datastore** within Azure Storage Ac

Azure Table stores **non-relational** structured data with a schemaless desig

There are two ways to interact with Azure Tables:

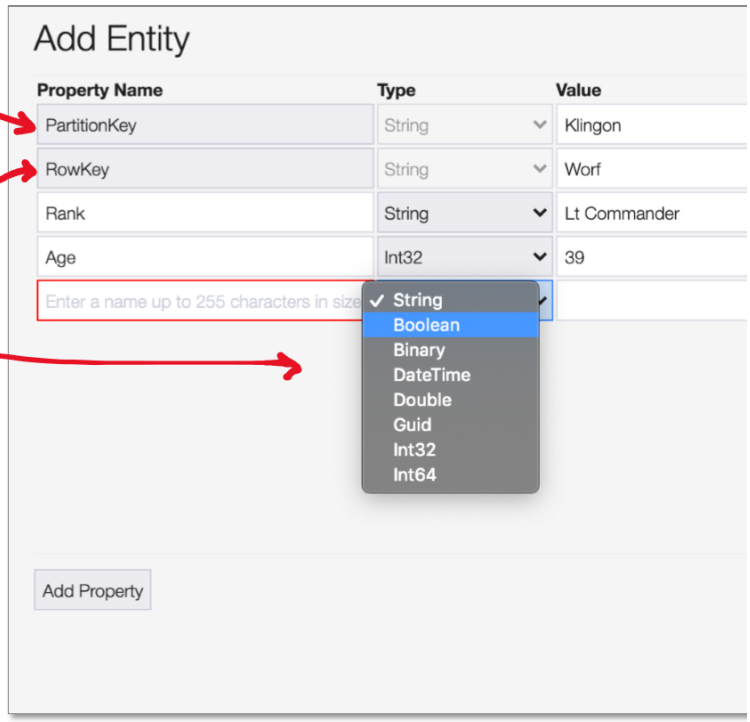
- Azure Table Storage API
- **Microsoft Azure Storage Explorer**



Azure Tables - Adding Entries

When you enter data you must provide a

- **Partition Key** — unique identifier for the partition within a given table
- **Row Key** — unique identifier for an entity within a given partition



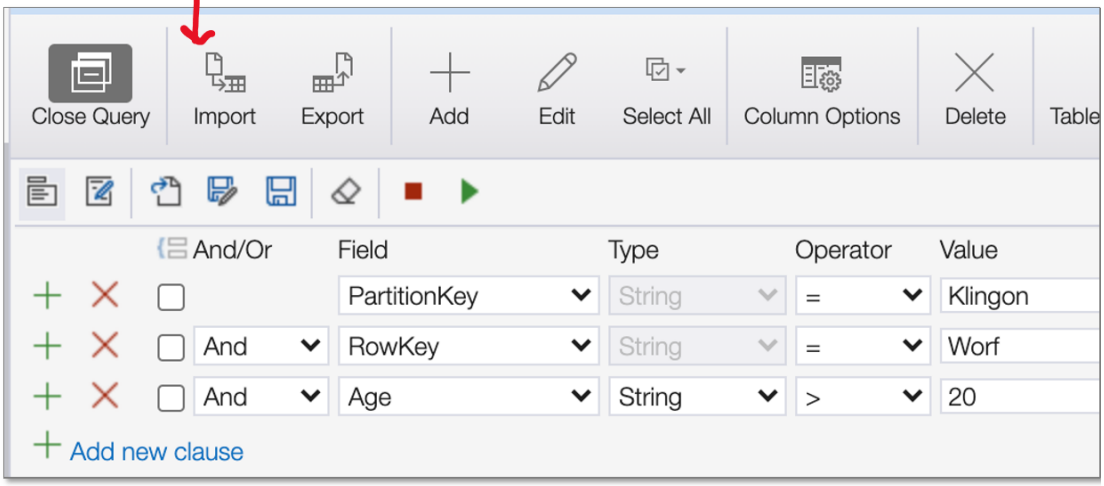
Data type

Azure Tables supports:

- String
- Boolean
- Binary
- DateTime
- Double
- Guid
- Int32
- Int64

Azure Tables – Querying

You can **query** data along the Partition and Row Key.
You also further filter for any filters



Introduction to CosmoDB

Azure CosmoDB is a service for fully-managed NoSQL databases that are designed to scale and high performance

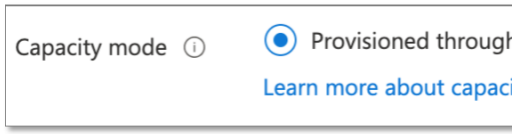
CosmoDB supports **different kinds** of NoSQL database engine which you intera

- **Core SQL** (**document** datastore)
- Azure Cosmos DB API for **MongoDB** (**document** datastore)
- **Azure Table** (**key/value** datastore)
- **Gremlin** (**graph** datastore)
  - Based on **Apache TinkerPop**

NoSQ

All of these NoSQL engines specific **capacity**:

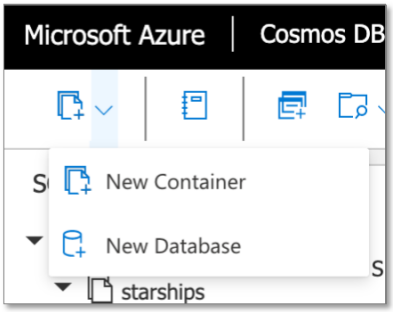
- Provisioned throughput (pay for guarantee of capacity)
- Serverless (pay for what you use)

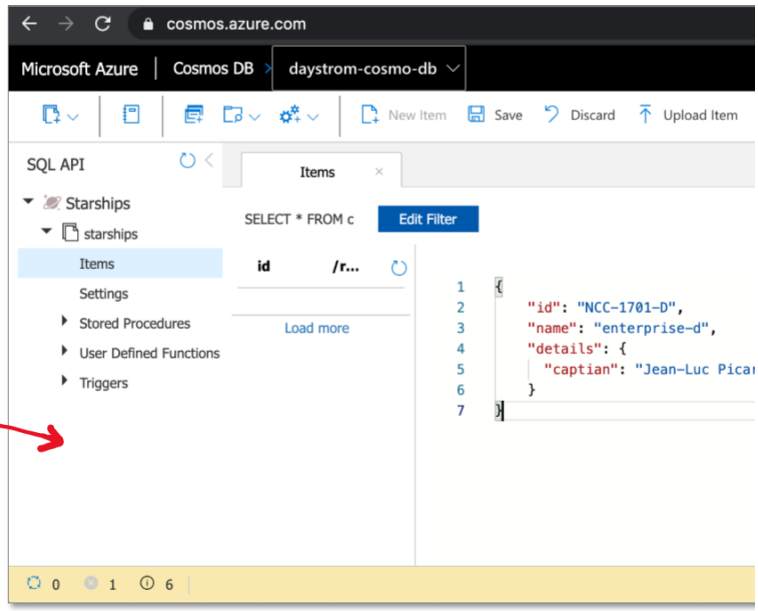


CosmoDB Explorer

CosmoDB Explorer is **a web interface** to explore and interact with your Cosmo

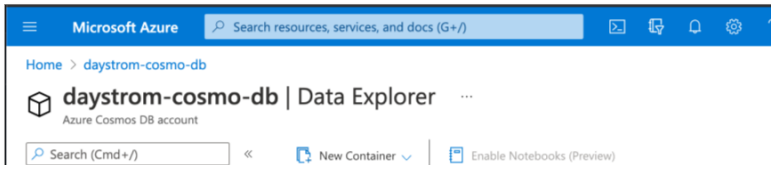
cosmos.azure.com



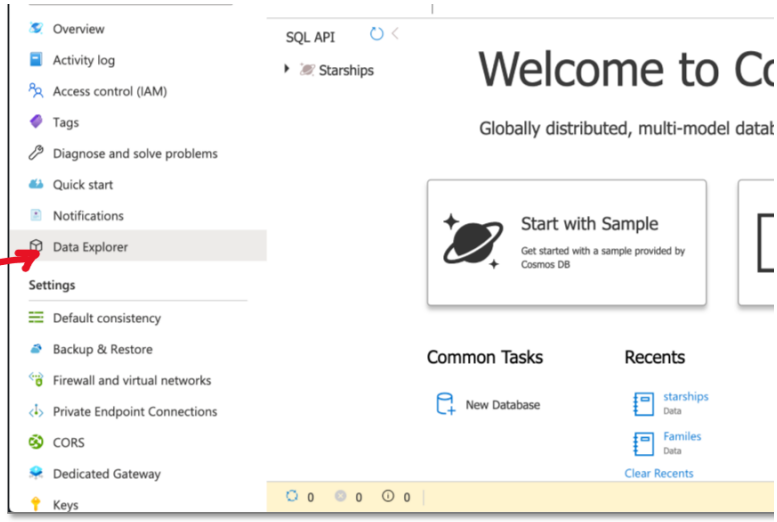


Here is a **CosmoDB Core SQL**
In CosmoDB Explorer adding a document to the database

CosmoDB Explorer

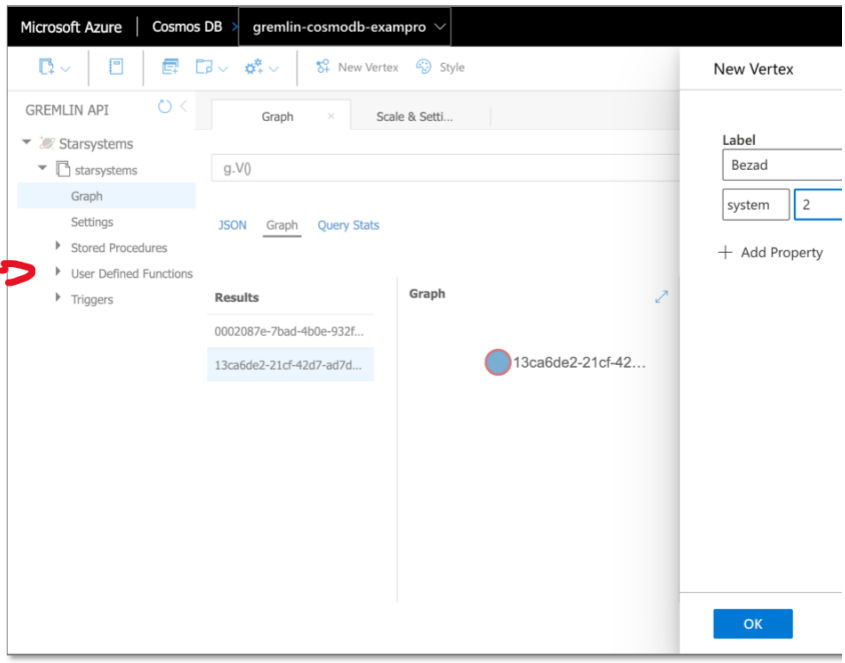


When navigating within Azure to a ComsoDB Account under **Data Explorer** is the same as CosmoDB Explorer



CosmoDB Explorer

Using CosmoDB Explorer With a **Graph Database** (Gremlin)



CosmoDB Table API vs Azure Table

When comparing Account Storage Azure Table vs CosmoDB Table API

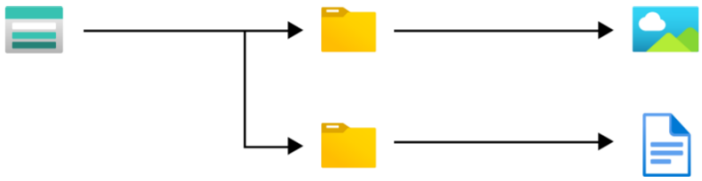
Feature	Azure Table Storage	Azure Cosmos DB Table API
Latency	Fast, but no upper bounds on latency.	Single-digit millisecond latency for reads
Throughput	Variable throughput model limit of 20,000 operations/s.	Guaranteed backed by SLAs. No upper limit on throughput
Global distribution	Single region with one optional readable	30+ regions

	secondary read region	
Indexing	Only primary index on PartitionKey and RowKey. No secondary indexes.	Automatic and complete indexing on all management.
Query	Query execution uses index for primary key, and scans otherwise.	Queries can take advantage of automati for fast query times.
Consistency	Strong within primary region Eventual within secondary region.	Five well-defined consistency levels
Pricing	Consumption-based	consumption-based or provisioned capa
SLAs	99.99% availability	99.99% availability SLA (some condition

Azure Blob

Blob storage is a **object-store** that is optimized for **storing massive amounts of unstructured**
Unstructured data is data that doesn't adhere to a particular data model or definition, such a

Azure Blobs are composed of the components:



Storage Account | Containers | Blobs

a unique namespace in Azure for your data
http://**mystorageaccount**.blob.core.windows.net

similar to a folder in a file system

The actual dat

Azure Blob

Azure Storage supports **3 types** of blobs:

1. **Block blobs**
   - store text and binary data
   - made up of blocks of data that can be managed individually
   - store up to about 4.75 TiB of data

2. **Append blobs**
   - Optimized for append operations

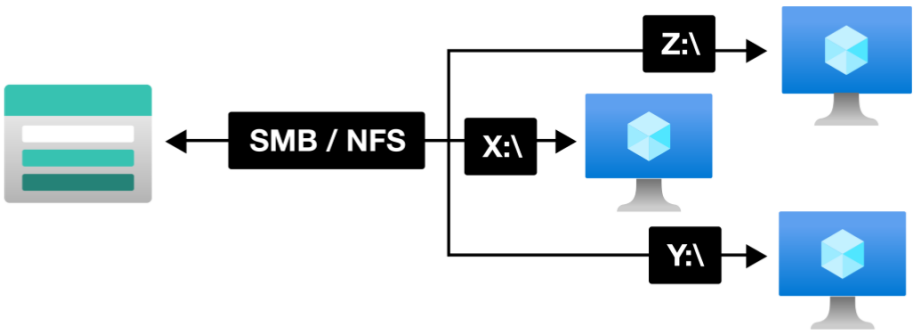
- ideal for scenarios such as logging data from virtual machine

3. Page blobs

- store random access files up to 8 TB in size.
- store virtual hard drive (VHD) files and serve as disks for Azure virtu

Azure Files

Azure Files is a fully managed **file share** in the cloud.
A file share is a **centralized server for storage** that allows **multiple connect**
It like having one big shared drive that everyone (Virtual Machines) can work on at t



To connect to the file share a **network protocol** is used:

- Server Message Block (SMB)
- Network File System (NFS)

When a connection is established the file share's filesystem will be accessible in the s
directory within your own directory tree. This is knowing as **mounting**

Azure Files – Use Cases

Use Cases

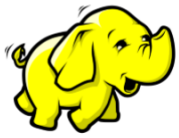
- Completely **replace or supplement** on-premises file servers Network Attach Storage (NAS) devices
- **Lift-and-Shift** your on-premise storage to the cloud via Classic Lift or Hybrid Lift
  - "Lift-and-Shift" means when you move workloads without rearchitecting, eg. importing local V
  - Classic Lift — where both the application and its data are moved to Azure
  - Hybrid Lift — where the application data is moved to Azure Files, and the application continue
- **Simplify cloud development**
  - Shared application settings — Multiple VMs and developer workstations need to access the sa
  - Diagnostic share — All VMs log to the file share, developers can mount and debug all logs in a
  - Dev/Test/Debug — Quickly share tools for developer needed for local environments
- **Containerization**
  - You can use Azure Files to persist volumes for stateful containers

Why use Azure files instead of setting up your own File Share server?

Why use Azure files instead of setting up your own File Share server?

- **Shared Access** — Already setup to work with standard networking protocols SMB and NFS
- **Fully managed** — Its kept up to date with security patches, designed to scale
- **Scripting and Tooling** — You can automate the management and creation of file shared with Azure
- **Resiliency** — Build to be durable and always working

What is Apache Hadoop?



Hadoop is an open-source framework for **distributed processing of large data sets**

Hadoop allows you to distribute:

- large dataset across many servers servers eg *HDFS*
- computing queries across many servers eg. *MapReduce*

These computer servers do not nee specialized hardware and can run o hardware

The Apache Hadoop framework has the following:

- Hadoop Common — collection of common utilities and libraries that support other Hadoop module
- Hadoop Distributed File System (HDFS) — a resilant and redudent file storage distrubted on clusters
- Hadoop MapReduce — writes apps that can process multi-terabyte data in-parallel on large clusters
- Hbase — a distributed, scalable, big data store
- YARN — manages resources, nodes, containers and performs scheduling
- HIVE — used for generating reports using an **SQL** language
- PIG — A high-level **scripting** language to write complex data transformations

Hadoop can integrate with many other open-source projects via **Hadoop com**

What is Apache Kafka?

Apache Kafka is an **open-source streaming platform** to create **high-perf pipelines**, streaming analytics, data integration, and mission-critical app

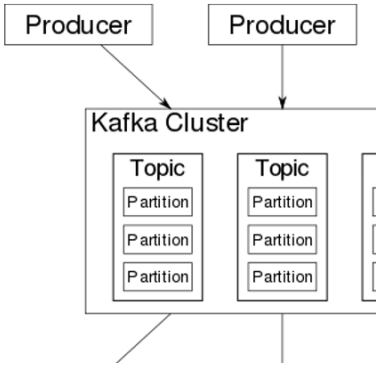
Kafka was originally developed by LinkedIn, and open-sourced in 2011

Kafka was written in Scala and Java.
to use Kafka you need to write **Java** code.

In Kafka data is stored in partitions on a Kafka Cluster which can span multiple machines (distributed computing)

Producers publish messages in a key and value format using the Kafka Producer API

Consumers can listen for messages and consume them using the Kafka Consumer API



Consumer Consumer

Messages are organized into **Topics**. Producers will push messages to topics and consumers will listen on topics.

Introduction to Azure HDInsights

Azure HDInsight is **managed service to run popular open-source analytics s**

HDInsights supports the following frameworks:

- Apache Hadoop
- Apache Spark
- Apache Kafka
- Apache Storm
- Apache Hive
- Apache HBase
- Low Latency Analytical Processing (LLAP)
- R

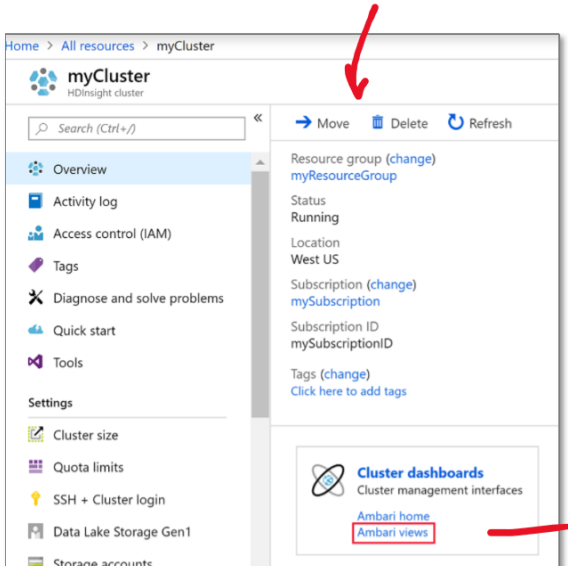
HDInsights has broad range of scenarios

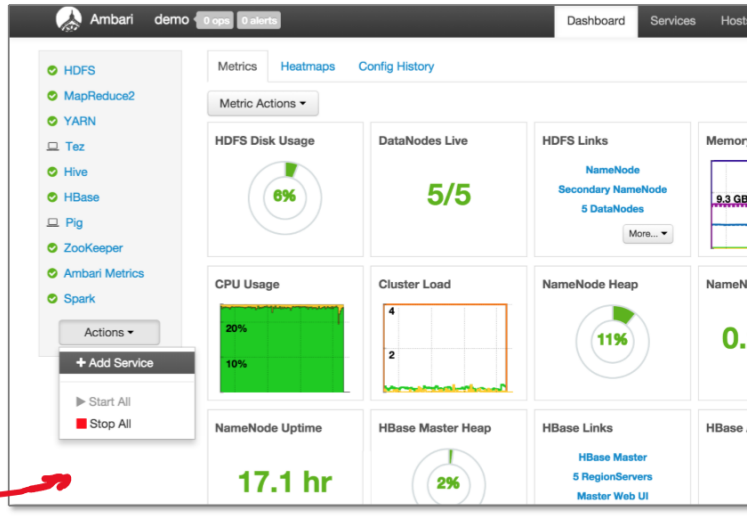
- Extract, Transform, and Load (ETL)
- Data Warehousing
- Machine Learning
- Internet of Things (IoT)

Azure HDInsights – Apache Ambari

Apache Ambari is an **open-source Hadoop management web-portal** for **provisioning**, **managing**, and **monitoring** Apache Hadoop clusters

When you **create an HDInsights Cluster** you will get a Cluster Dashboard (Apache Ambari dash





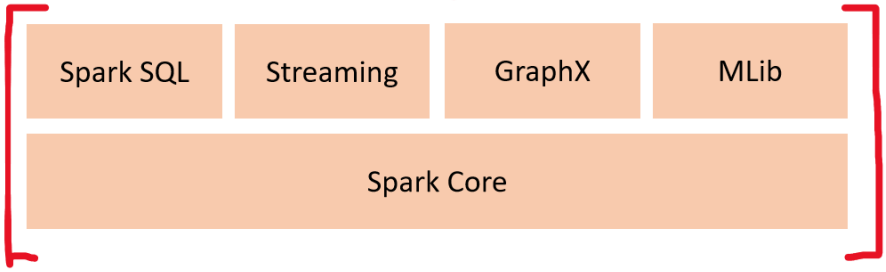
Apache Spark

Apache Spark is an open-source **unified analytics engi**
big data and machine learning

Spark lets you run workloads much faster than Hadoop:
- 100x faster in memory
- 10x faster than disk

Which is why Spark is described as ***lightning fast***

Apache Spark is a collection of libraries that work well together to form an **analytics ecosystem**



Spark Core
The underlying engine and API. The Al
programming languages:
R, SQL, Python, Scala, Java

Spark SQL
Introduces a data structure called a Da
used with DSL to work with structure

Spark Streaming
Allows Spark to ingest data from many
HDFS, Flume, Kafka, Twitter, Kinesis

GraphX
distributed graph-processing framewo

Machine Learning Library (MLib)
a distributed machine-learning framev
machine learning and statistical algori

Apache Spark – RDD API

To interact with Apache Spark:

Resilient Distributed Dataset (RDD) is a domain specific language (DSL) to execute various parall
operations on an Apache Spark cluster.

Common RDD API functions

map
flatMap
mapPartitions
filter
distinct
reduce
count
first
take
countByValue
sortBy
groupBy
fold

union
add
substract
intersection
saveAsTextFile
saveAsHadoopFile
saveAsPicketFile
min
max
mean
status
parallelize

Example of RDD API

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.
             .map(lambda word: (word, 1)) \
             .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```

Databricks Platform

Databricks is a software company specializing in **providing fully managed Apac**
The company founders were the creators of Apache Spark, Delta Lake and MLF

Databricks has two offerings:

Databricks Platform — Databricks cloud-based Spark platform with an ease-to-
- Launch fully managed Spark clusters
- Launch notebooks to write code and interact with Spark
- Create workspaces to collaborate with team members
- Role Base Access Controls
- Create jobs for ELT or data analysis tasks that run immediately or on a schedu
- Create MLFlow Workflows
- **Available on all main cloud service providers eg. AWS, Azure, GCP**

free

Databricks Community Edition – free version of Databricks Platform for educati
- Create a free micro-cluster that terminates after 2 hours when idle
- No workspace, jobs or RBAC

Role Base Access Control

Introduction to Azure Databricks

Azure Databricks is a **partnership between Microsoft and Databricks** to offe
Databricks Platform within the Azure Portal running on Azure compute serv

Azure Databricks offers two environments:

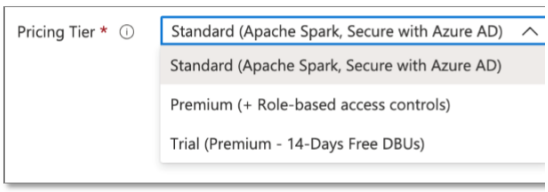
Azure Databricks Workspace
- The DataBrick Platform with integrations to **Azure data-related services** for building big data pi
  - Batching: Azure Data Factory
  - Streaming: Apache Kafka Event Hub, or IoT Hub
  - Storage: Azure Blob Storage or Azure Data Lake Storage

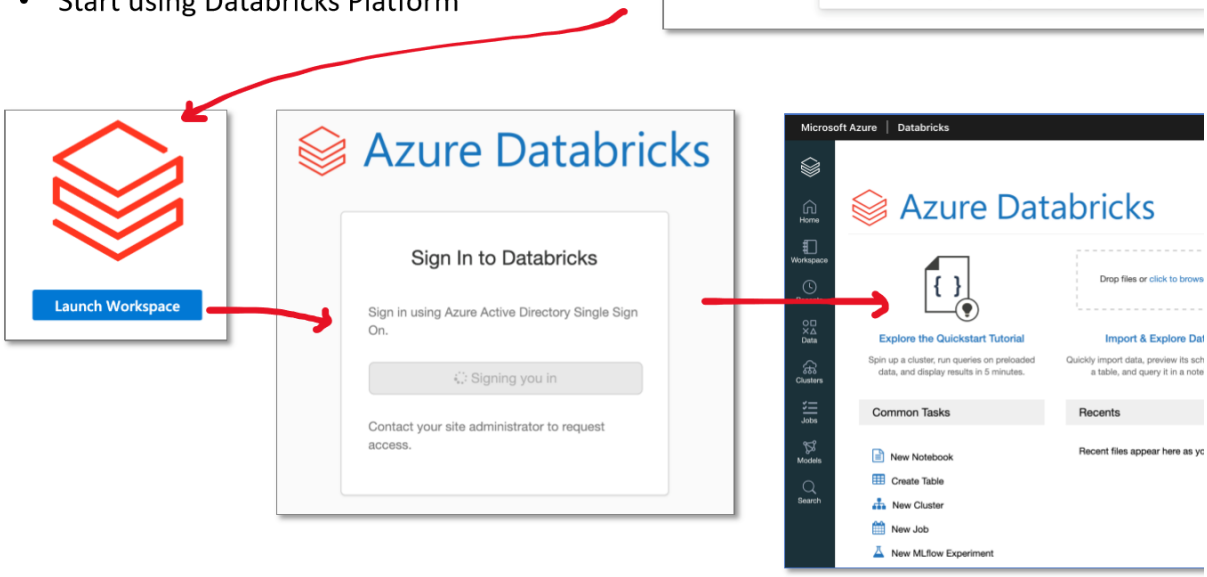
Azure Databricks SQL Analytics
- run SQL queries on your data lake
- create multiple visualization types to explore query results
- build and share dashboards

Azure Databricks Workspaces

Create a Databricks Workspace by:

- Creating a workspace and choosing a plan
- Launching the workspace
- SSO to the the workspace
- Start using Databricks Platform



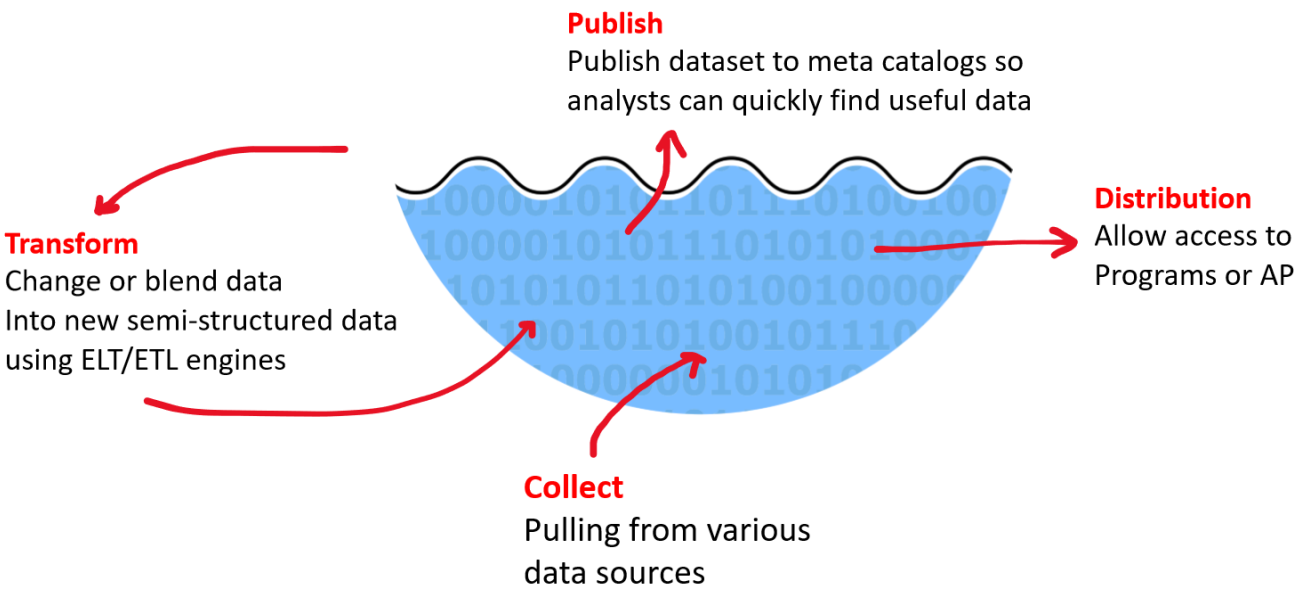


Introduction to Data Lakes

A data lake is a **centralized data repository for unstructured and semi-structu**

A Data Lake is intended to store vast amounts of data

Data lakes generally use **object (blobs) or files** as its storage medium.

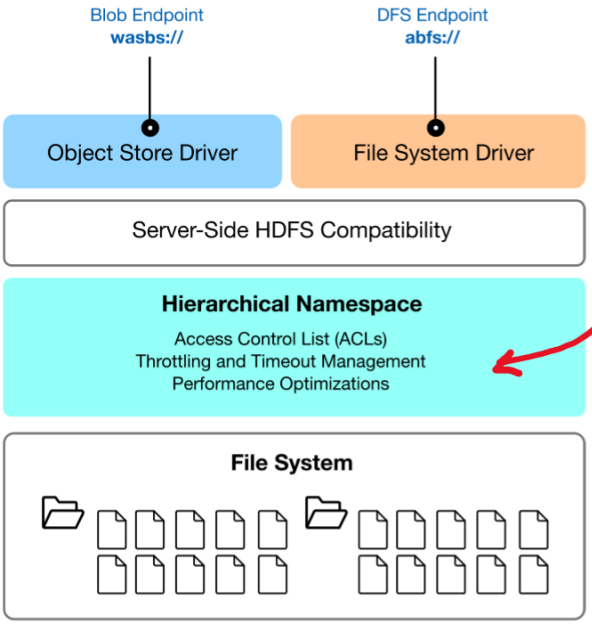


Introduction to Azure Data Lake

Azure Data Lake Storage Gen1 *(deprecated)*

The first version of Data Lake Storage and will be retired in 2024. New users should use Gen2

The first version of Data Lake Storage and will be retired in 2024. New users should use Gen2



Azure Data Lake Storage Gen2

Data Lake Storage is Azure Blog storage which is has l
extended to support big data analytics workloads

- Designed to handle **petabytes of data** and **hundre gigabits of throughput**
- In order to efficiently access data, Data Lake Stora **hierarchical namespace** to Azure Blob Storage

Azure Data Lake Storage Gen2

Azure Data Lake Storage can be accessed two ways:

Azure Blob File System over SSL (abfss)

An Hadoop filesystem driver that is compatible with Azure Data Lake Storage Gen2
The abbfss driver will have a URL that will allow you to access and write to the underlying Data Lake S

```
abfs[s]://<file_system>@<account_name>.dfs.core.windows.net/<path>/<fi
```

Windows Azure Storage Blob over SSL (wasbs)

<MAYBE JUST SKIP THIS SLIDE>

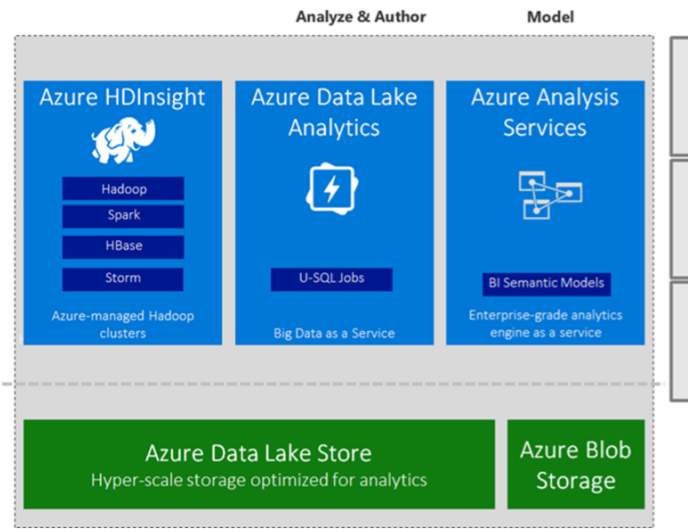
Azure Data Lake Analytics

Azure Data Lake Analytics is an **on-demand analytics job service** that simplifies

Instead of deploying, configuring, and tuning hardware…

You write queries (via U-SQL) to transform your data and extract valuable insig

Exporting approximately 2.8 billion rows of TPC-DS store sales data (~500 GB) into a CSV file took less than 7 minutes and importing the full 1 TB set of source data into Azure Analysis Services by using the Azure Data Lake connector took less than 6 hour



Azure Data Lake Analytics – U-SQ

U-SQL is a structured query language included within Data Lake Analytics to **perform queries** on your data lake.

```
DECLARE @in  string = "/Samples/Da
DECLARE @out string = "/output/res

@searchlog =
    EXTRACT UserId          int,
            Start           DateTi
            Region          string
            Query           string
            Duration        int?,
            Urls            string
            ClickedUrls     string
    FROM @in
    USING Extractors.Tsv();

@rs1 =
    SELECT Start, Region, Duration
    FROM @searchlog
    WHERE Region == "en-gb";

@rs1 =
    SELECT Start, Region, Duration
    FROM @rs1
    WHERE Start >= DateTime.Parse(
          AND Start <= DateTime.Pa

OUTPUT @rs1
TO @out
```

U-SQL can query and combine data from a variety of data sources, including:

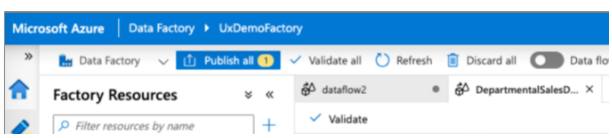
- Azure Data Lake Storage
- Azure Blob Storage
- Azure SQL DB
- Azure SQL Data Warehouse,
- SQL Server instances running in Azure VMs

You can install **Azure Data Lake Tools** for Visual Studio to perform U-SQL jobs on your Azure Data Lake

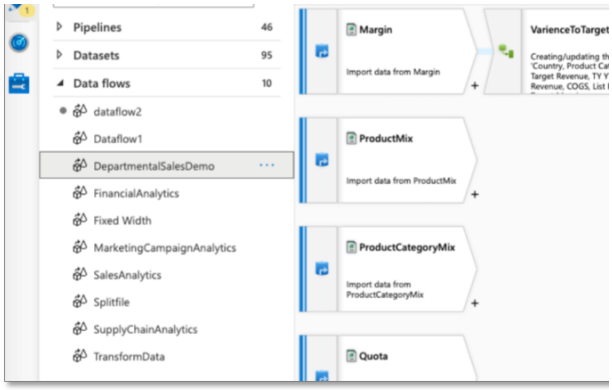
Introduction to Azure Data Factory

Azure Data Factory is a managed service for **ETL, ELT and data integ**
Create **data-driven workflows** for orchestrating **data movement** and **transform**

- Create Pipelines to schedule data-driven workflows



- build complex ETL processes that transform data visually with data flows
- using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database
- publish your transformed data to data stores such as Azure Synapse Analytics
- raw data can be organized into meaningful data stores and data lakes



Azure Data Factory – Core Componer

Pipelines
a logical grouping of activities that performs a unit of work

Activities
A processing step in a pipeline

Datasets
data structures within the data store

Linked services
define the connection information for data sources to connect to Data Factory

Data Flows
logic to determine how data moves through a pipeline or is transformed

Integration Runtimes (RI)
compute infrastructure used by Azure Data Factory

Control flow
orchestration of pipeline activities that includes chaining activities in a sequence, branching

Microsoft SQL Server Integration Serv

Microsoft SQL Server Integration Services

Microsoft SQL Server Integration Services (SSIS) is a platfor
enterprise-level data integration and data transformatio

You can perform the following tasks with SSIS
- Copy files
- Download files
- Loading data into data warehouses
- Cleansing data
- Mining data

SSIS can be used to automate SQL Server databas
SSIS can be used as an integration runtime in Azu

SSIS has...
- built-in tasks and transformations

- managing SQL Server objects
- managing SQL Server data

Perform ELT with variety of sources:
- XML
- Flat files
- Relational data sources

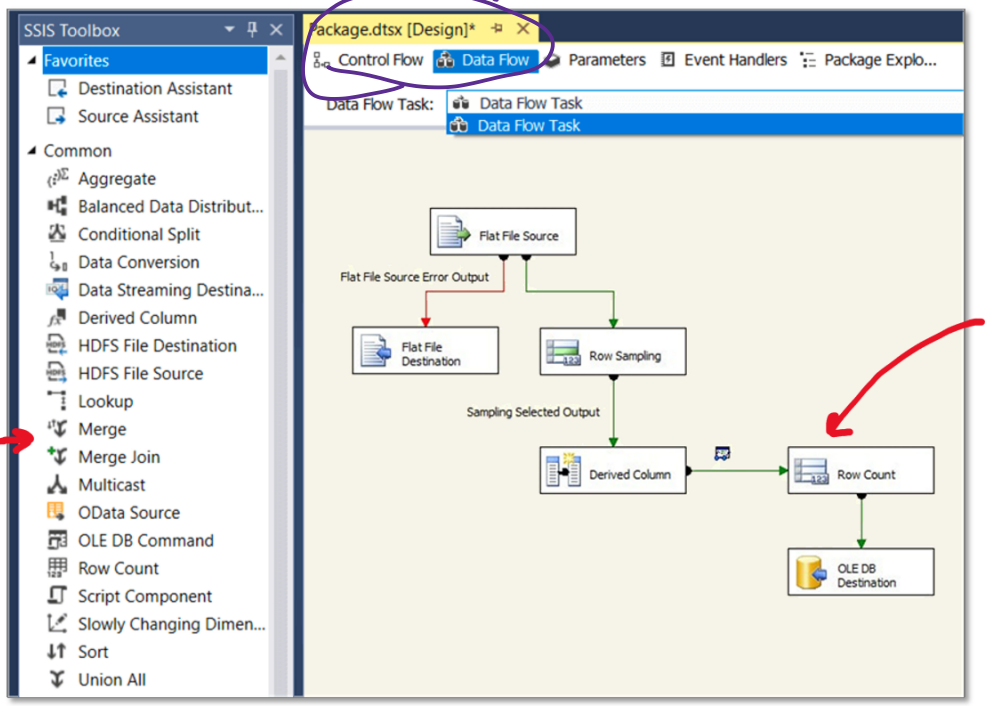
- graphical tools for building packages
- Integration Services Catalog database
  - where you store, run, and manage packages

Use **Graphical Integration Services tools** for to integr and transform data without having to write code

SSIS Designer is a graphical tool that you can use to maintain Integration Services packages.

Microsoft SQL Server Integration Serv

similary to Data fac



SSIS allows you to drag out data transformations