Design and implement data storage – Basics

Understanding Data

Data storage **Data processing** **Visualizing your data**

Classification of data

Structured data - Tabular data that is represented by rows and columns in a database

	ProductID	Name	ProductNumber	Color	StandardCost	ListPrice	Size	Weight	ProductCategoryID	ProductModelID
1	680	HL Road Frame - Black, 58	FR-R92B-58	Black	1059.31	1431.50	58	1016.04	18	6
2	706	HL Road Frame - Red, 58	FR-R92R-58	Red	1059.31	1431.50	58	1016.04	18	6
3	707	Sport-100 Helmet, Red	HL-U509-R	Red	13.0863	34.99	NULL	NULL	35	33
4	708	Sport-100 Helmet, Black	HL-U509	Black	13.0863	34.99	NULL	NULL	35	33
5	709	Mountain Bike Socks, M	SO-B909-M	White	3.3963	9.50	M	NULL	27	18
6	710	Mountain Bike Socks, L	SO-B909-L	White	3.3963	9.50	L	NULL	27	18
7	711	Sport-100 Helmet, Blue	HL-U509-B	Blue	13.0863	34.99	NULL	NULL	35	33
8	712	AWC Logo Cap	CA-1098	Multi	6.9223	8.99	NULL	NULL	23	2
9	713	Long-Sleeve Logo Jersey, S	LJ-0192-S	Multi	38.4923	49.99	S	NULL	25	11
10	714	Long-Sleeve Logo Jersey, M	LJ-0192-M	Multi	38.4923	49.99	M	NULL	25	11
11	715	Long-Sleeve Logo Jersey, L	LJ-0192-L	Multi	38.4923	49.99	L	NULL	25	11
12	716	Long-Sleeve Logo Jersey, XL	LJ-0192-X	Multi	38.4923	49.99	XL	NULL	25	11
13	717	HL Road Frame - Red, 62	FR-R92R-62	Red	868.6342	1431.50	62	1043.26	18	6
14	718	HL Road Frame - Red, 44	FR-R92R-44	Red	868.6342	1431.50	44	961.61	18	6
15	719	HL Road Frame - Red, 48	FR-R92R-48	Red	868.6342	1431.50	48	979.75	18	6
16	720	HL Road Frame - Red, 52	FR-R92R-52	Red	868.6342	1431.50	52	997.90	18	6
17	721	HL Road Frame - Red, 56	FR-R92R-56	Red	868.6342	1431.50	56	1016.04	18	6
18	722	LL Road Frame - Black, 58	FR-R38B-58	Black	204.6251	337.22	58	1115.83	18	9

The database engine is responsible for storage of data

When you query for data , it is responsible for giving you the results

Install a database engine on a virtual machine

Oracle, Microsoft SQL Server, MySQL

Query - select * from Product where Color = 'Black'

Semi-structed data

This is data that resides in other formats and not in a database as such

Common example - JSON - JavaScript Object Notation

```
{
  "customerid": 1,
  "customername" : "John",
  "city" : "Miami"
}
```

key-value data store

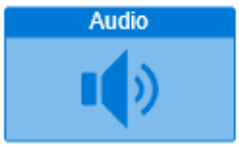
Here you get the value based on the key that is being stored

customerid	customername	city
1	John	Miami
2	Clark	Chicago
3	Emily	New York
4	Sarah	Miami

key-value data store can be stored in the form of tables

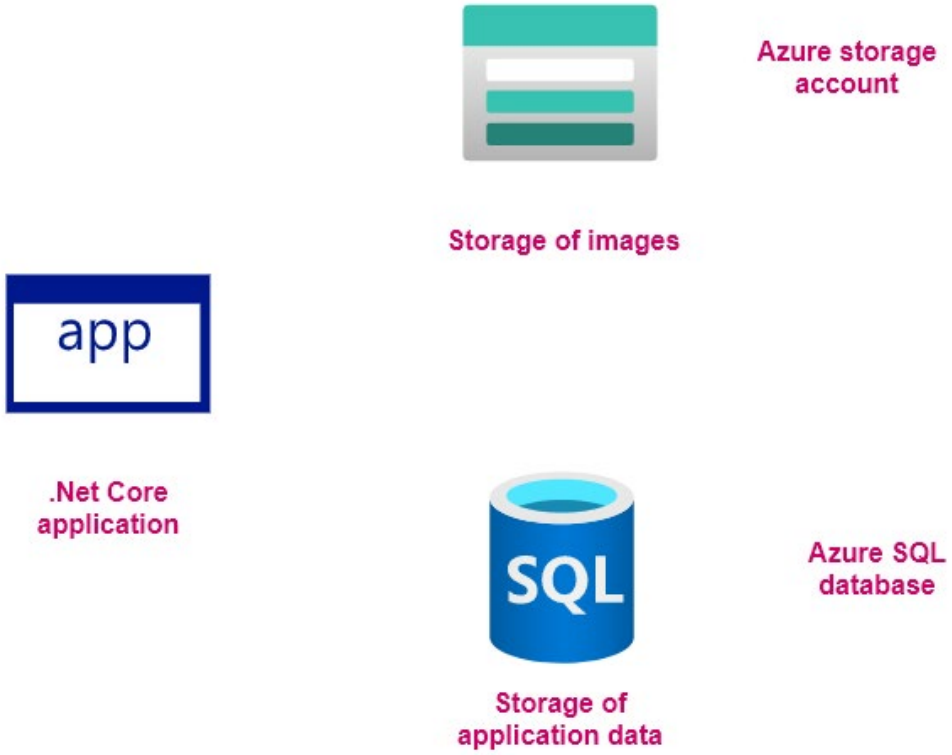
But here the data does not have a specific structure or relation

unstructured data

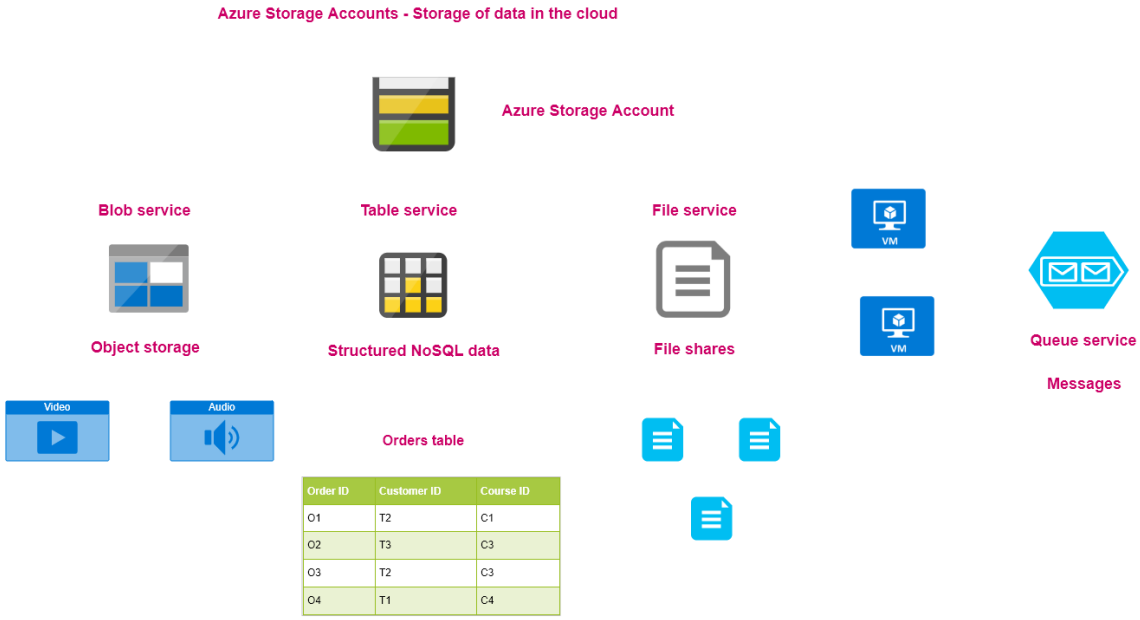


These are all binary objects

Example of data storage



Lab - Azure Storage accounts



Lab - Azure SQL databases

Host a SQL database

1. Create a virtual machine

2. Install the database software

3. Create your database , tables and store your data

4. Administrative tasks - Backup, High Availability

Azure SQL database

This is a platform as a service

Here the infrastructure is managed for you

Lab - Application connecting to Azure Storage and SQL database

Azure Data Lake Storage Gen2

This service is built on top of Azure Blob storage

Gives the ability to host an enterprise data lake on Azure

You also get the feature of a hierarchical namespace on top of Azure Blob storage

Helps to organize objects/files into a hierarchy of directories for efficient data access

A data lake is used to store large amounts of data in its native, raw format

Data lakes are optimized for storing terabytes and petabytes of data

The data could come from a variety of data sources

The data itself could be in various formats - Structured, semi-structured and unstructured data

Different file formats

JSON

JavaScript Object Notation


```
{
  "count": 1,
  "total": 0,
  "minimum": 0,
  "maximum": 0,
  "average": 0,
  "resourceId": "/SUBSCRIPTIONS/E5250E15-0516-48F0-889B-DAE6C15B6529
    /RESOURCEGROUPS/PRODUCTIONGRP/PROVIDERS/MICROSOFT.DBFORMYSQL
    /SERVERS/WORDPRESS-SERVER2020",
  "time": "2021-02-16T17:36:00.0000000Z",
  "metricName": "cpu_percent",
  "timeGrain": "PT1M"
}
```


The JSON contents are enclosed in curly brackets. It is a JSON document

Each document consists of fields. Each field has a name and value

Avro

This is a row-based format file

Each record in the file contains a header that describes the structure of the data in the record

The data iself is stored in binary format

This format is ideal for compressing data

Results in less storage

Requires less bandwith requirements

Parquet data format

This is a columnar data format

It was created by Cloudera and Twitter

Data for each column is stored together in something known as a row group

This data format support compression and different encoding schemes

Azure Storage Account – Redundancy

Azure Storage account - Redundancy

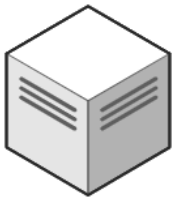
Multiple copies of your data are stored

This helps to protect against planned and unplanned events - transient hardware failures, network or power outages.

Storage Device

Locally-redundant storage

Data Center



Central US

Here three copies of your data are made

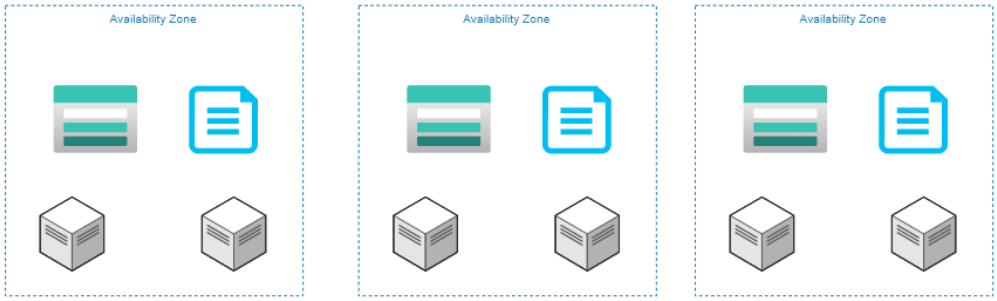
It helps to protect against server rack of drive failures

Storage Device Storage Device Storage Device

Zone-redundant storage

This helps to protect against data center level failures

Here data is replicated synchronously across three Azure availability zones

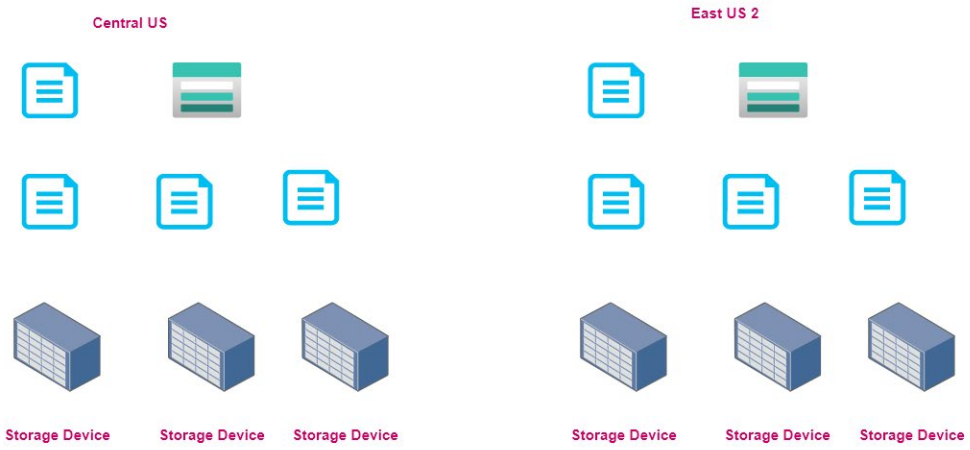


Central US

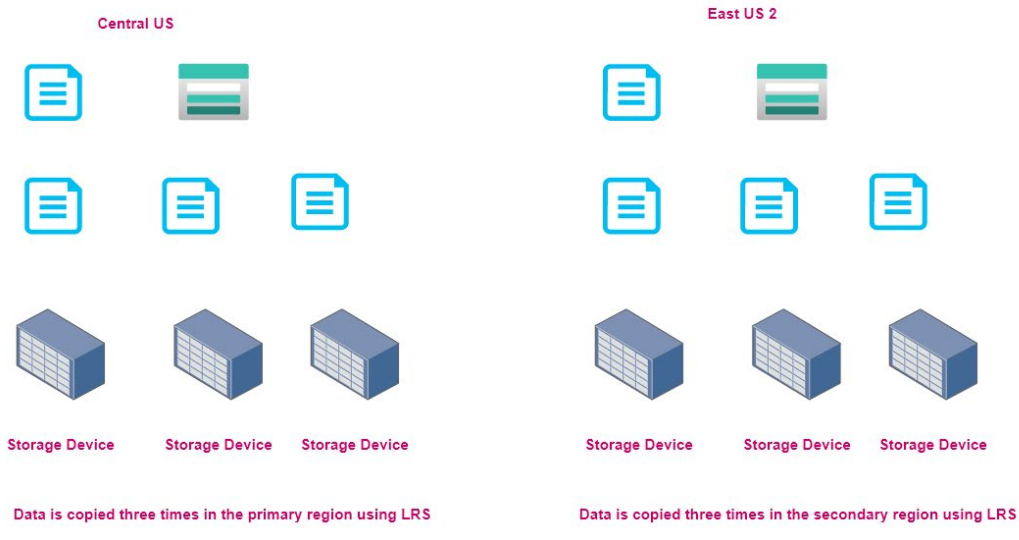
Each availability zone is a seperate physical location with independent power, cooling and networking

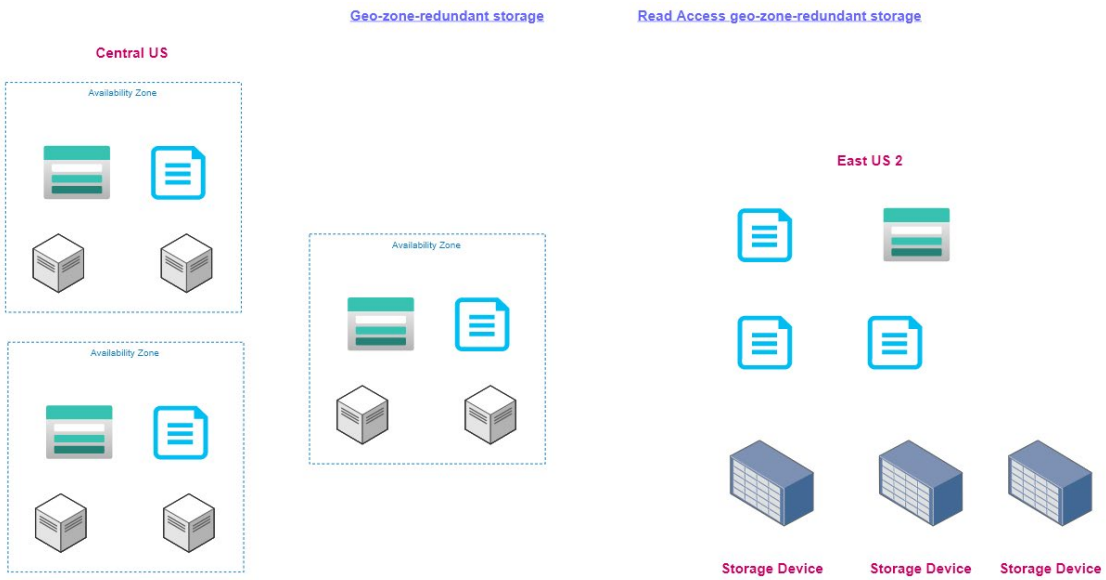
Geo-redundant storage

Here data is replicated to another region



Read-access geo-redundant storage



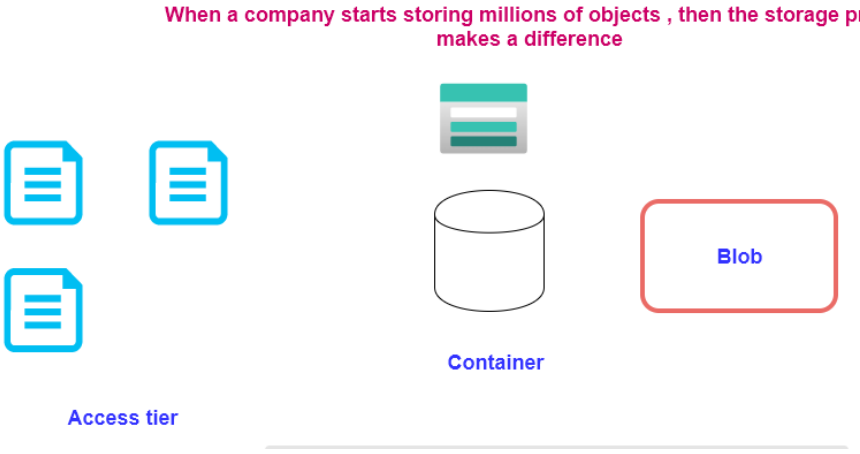


Azure Storage Account - Access tiers

Data storage prices pay-as-you-go

All prices are per GB per month.

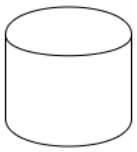
	PREMIUM	HOT	COOL	ARCHIVE
First 50 terabyte (TB) / month	$0.15 per GB	$0.0184 per GB	$0.01 per GB	$0.00099 per GB
Next 450 TB / month	$0.15 per GB	$0.0177 per GB	$0.01 per GB	$0.00099 per GB
Over 500 TB / month	$0.15 per GB	$0.0170 per GB	$0.01 per GB	$0.00099 per GB



Hot Cool Archive

The Archive can only be enabled at the individual blob level

Archive



Container

You have to rehydrate the file to access the file

Here you need to change the access tier of the file to either Hot or Cool to access the file

It takes time to rehydrate the file

	PREMIUM	HOT	COOL	ARCHIVE
Write operations (per 10,000)[1]	$0.0175	$0.05	$0.10	$0.10
List and Create Container Operations (per 10,000)[2]	$0.05	$0.05	$0.05	$0.05
Read operations (per 10,000)[3]	$0.0014	$0.004	$0.01	$5
Archive High Priority Read (per 10,000)[5]				$50
All other Operations (per 10,000), except Delete, which is free	$0.0014	$0.004	$0.004	$0.004

Early Deletion Fee

Cool Access tier - This is used for data that is accessed infrequently and stored for at least 30 days

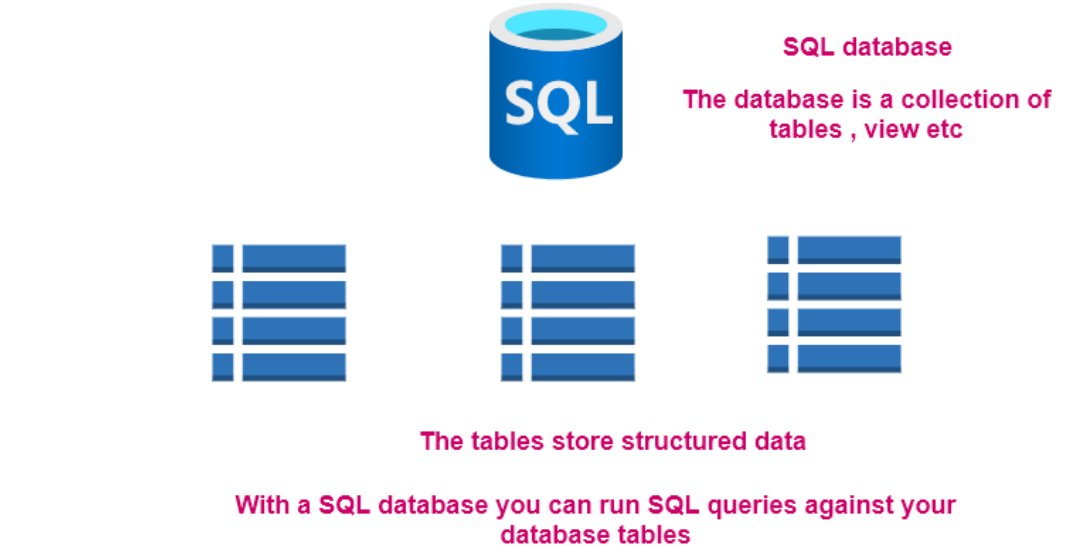
Archive Access tier - This is used for data that is rarely accessed and stored for at least 180 days

If you have a blob in the Cool Access tier and you change the access tier to the Hot access tier earlier than 30 days , then you are charged an early deletion fee

If you have a blob in the Cool Access tier and you change the access tier to the Hot access tier after just 10 days, then you are still charged costs for the extra 20 days of the Cool Access tier

Design and implement data storage - Overview on Transact-SQL

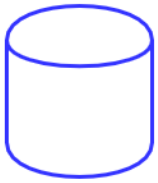
The internals of a database engine



Design and implement data storage - Azure Synapse Analytics

Why do we need a data warehouse

The need for a SQL data warehouse



Transactional system

Here the system will mostly handle millions of transactions on a daily basis

Online Transactional Processing system

	ProductID	Name	ProductNumber	Color	StandardCost	ListPrice	Size	Weight	ProductCategoryID	ProductModelID
1	680	HL Road Frame - Black, 58	FR-R92B-58	Black	1059.31	1431.50	58	1016.04	18	6
2	706	HL Road Frame - Red, 58	FR-R92R-58	Red	1059.31	1431.50	58	1016.04	18	6
3	707	Sport-100 Helmet, Red	HL-U509-R	Red	13.0863	34.99	NULL	NULL	35	33
4	708	Sport-100 Helmet, Black	HL-U509	Black	13.0863	34.99	NULL	NULL	35	33
5	709	Mountain Bike Socks, M	SO-B909-M	White	3.3963	9.50	M	NULL	27	18
6	710	Mountain Bike Socks, L	SO-B909-L	White	3.3963	9.50	L	NULL	27	18
7	711	Sport-100 Helmet, Blue	HL-U509-B	Blue	13.0863	34.99	NULL	NULL	35	33
8	712	AWC Logo Cap	CA-1098	Multi	6.9223	8.99	NULL	NULL	23	2
9	713	Long-Sleeve Logo Jersey, S	LJ-0192-S	Multi	38.4923	49.99	S	NULL	25	11
10	714	Long-Sleeve Logo Jersey, M	LJ-0192-M	Multi	38.4923	49.99	M	NULL	25	11
11	715	Long-Sleeve Logo Jersey, L	LJ-0192-L	Multi	38.4923	49.99	L	NULL	25	11
12	716	Long-Sleeve Logo Jersey, XL	LJ-0192-X	Multi	38.4923	49.99	XL	NULL	25	11
13	717	HL Road Frame - Red, 62	FR-R92R-62	Red	868.6342	1431.50	62	1043.26	18	6
14	718	HL Road Frame - Red, 44	FR-R92R-44	Red	868.6342	1431.50	44	961.61	18	6
15	719	HL Road Frame - Red, 48	FR-R92R-48	Red	868.6342	1431.50	48	979.75	18	6
16	720	HL Road Frame - Red, 52	FR-R92R-52	Red	868.6342	1431.50	52	997.90	18	6
17	721	HL Road Frame - Red, 56	FR-R92R-56	Red	868.6342	1431.50	56	1016.04	18	6
18	722	LL Road Frame - Black, 58	FR-R38B-58	Black	204.6251	337.22	58	1115.83	18	9

Udemy

Register as a student

Table

Search for a course

Table

Purchase the product

Table

Online Analytical Processing system

Here analytics is performed on the underlying data

Helps business users get a better understanding on the data

Register as a student

1. How many students are registering per day?

2. Which countries have the most number of student registrations?

Search for a course

1. How many courses are being added by instructors per day?

2. Which courses are being searched more?

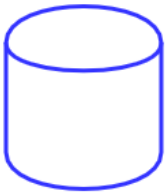
Course Table

Purchase the product

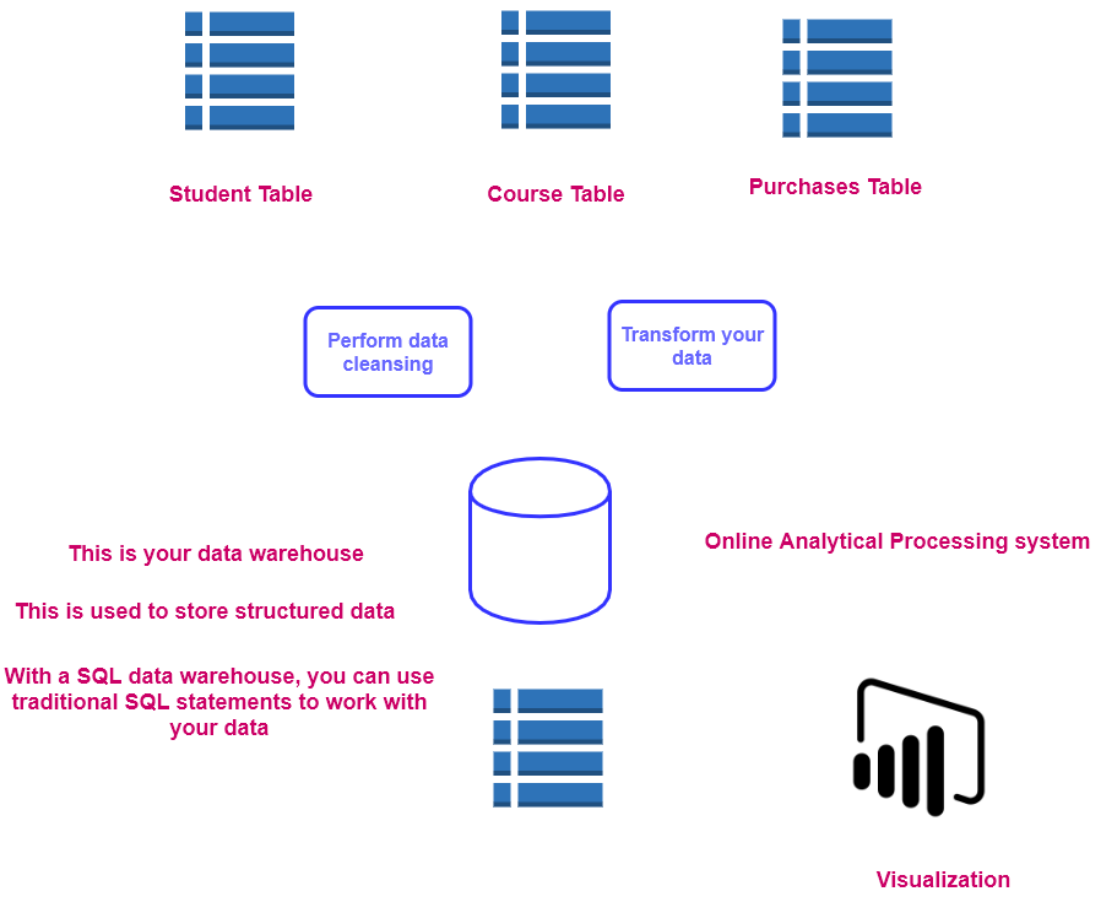
1. How many courses are being purchased per minute?

2. Which are the most popular courses being purchased?

Purchases Table

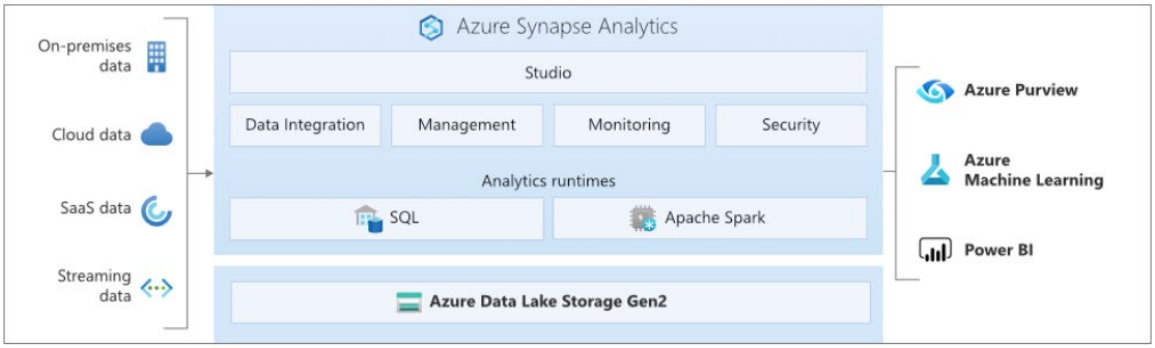


Online Transactional Processing system



Welcome to Azure Synapse Analytics

This is an enterprise analytics service



Here you can host your data warehouse with the use of SQL pools

You can use Pipelines for data integration. This allows you to perform ETL/ELT activities to being the data into your data warehouse

You can also use the power of Spark for processing your data.

Helps to bring you data lakes closer to you and process them with integration with Azure storage - Azure Data lake Storage accounts

Integrates with Azure services like Azure Monitor and Azure Active Directory.

Azure Synapse - Compute options

There are 2 compute options
we are going to look at

Serverless SQL pool	SQL Pool
You can use this option to perform quick adhoc analysis of data	This is used to build your data warehouse
You can use T-SQL to work with your data	If you need to persist your data
Here you are charged based on how much you use	Here compute nodes will be used to process the data
This is because there is no underlying infrastructure	Here you are charged based on a metric known as DWU
The charge is based on how much data you process	

Using External tables

We are going to define external tables

Here the data lies in another source and we just defined the table structure in Azure Synapse

When we query for data within the table, the data is queried from the external source

An external table can point to data that is located in Hadoop, Azure Blob storage or Azure Data Lake Storage

Azure Synapse

Serverless pool

PolyBase

Azure Data Lake Storage

Container - data

CSV file

Step 1 : Authorization to use the Data Lake storage account

Step 2 : Define the format of the external file we are going to work with - Parquet, CSV

Step 3 : Create the external table

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=native

External table type	Hadoop	Native
Dedicated SQL pool	Available	Parquet tables are available in **gated preview** - contact your Microsoft Technical Account Manager or Cloud Solution Architect to check if you can add your dedicated SQL pool to the gated preview.
Serverless SQL pool	Not available	Available
Supported formats	Delimited/CSV, Parquet, ORC, Hive RC, and RC	Serverless SQL pool: Delimited/CSV, Parquet, and Delta Lake(preview) Dedicated SQL pool: Parquet
Folder partition elimination	No	Only for partitioned tables synchronized from Apache Spark pools in Synapse workspace to serverless SQL pools
Custom format for location	Yes	Yes, using wildcards like `/year=*/month=*/day=*`
Recursive folder scan	No	Only in serverless SQL pools when specified `/**` at the end of the location path
Storage filter pushdown	No	Yes in serverless SQL pool. For the string pushdown, you need to use `Latin1_General_100_BIN2_UTF8` collation on the `VARCHAR` columns.
Storage authentication	Storage Access Key(SAK), AAD passthrough, Managed identity, Custom application Azure AD identity	Shared Access Signature(SAS), AAD passthrough, Managed identity

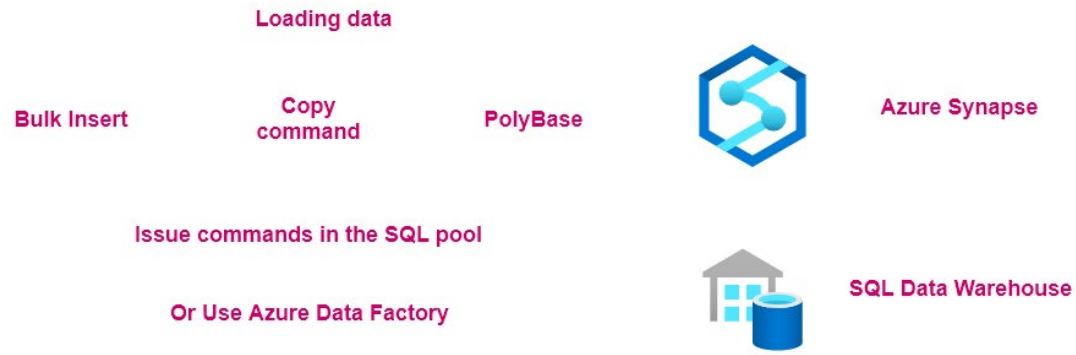
Loading data into the Dedicated SQL Pool

Some times data needs to be loaded and available in real-time

Other times data might be loaded on a daily basis at the end of the day.
Here you just perform incremental loading of the data

Ways of loading data

Using the Copy statement	Using Transact-SQL , you can transfer data into a table in a SQL pool
Azure Synapse pipeline	Here you can also perform transformations on your data
Using Polybase to define external tables	Here your data can be in an external data store. But you can access the data via external tables



Using Polybase is the best choice for loading or exporting large volumes of data and when you need fast performance whilst loading data

Polybase only supports loading files which are available in Azure Blob storage and Azure Data Lake Storage

Data Cleansing

Data cleansing or data cleaning

This is the process of finding and correcting/removing corrupt or inaccurate records in a record set

What do you do with rows that have columns with NULL values

What do you do with records that have duplicate row values

Sometimes values may not be in the defined range - For example , the age of a person

Formating dates, different systems might store dates in different formats

Designing a data warehouse

Fact Tables

These are measurements or metrics that correspond to facts

For example - Sales Table - This records all the sales that have been made

The sales data are facts that sales have actually been made

Dimension tables

This helps to provide some sort of context to the facts that are presented

For example - What are the products that were sold

Who are the customers who bought the products

Building your fact table

Building facts around the sales data

The sales data will keep on getting added to the sales fact table

OLTP
Transactional data

OLAP
Analysis

Orders/Sales data

This is your fact table

It contains facts about the sales made for your courses

Order ID	Course ID	quantity
O1	C1	10
O2	C2	20
O3	C3	30
O4	C3	40

Sales table

Transactions for 2016

Transactions for 2017

Transactions for 2018

Transactions for 2019

Transactions for 2020

Your Dimension Table

OLTP
Transactional data

OLAP
Analysis

Courses
AZ-900
DP-203
AZ-104

Customer
UserA
UserB
UserC

Instructor
InstructorA
InstructorB
InstructorC

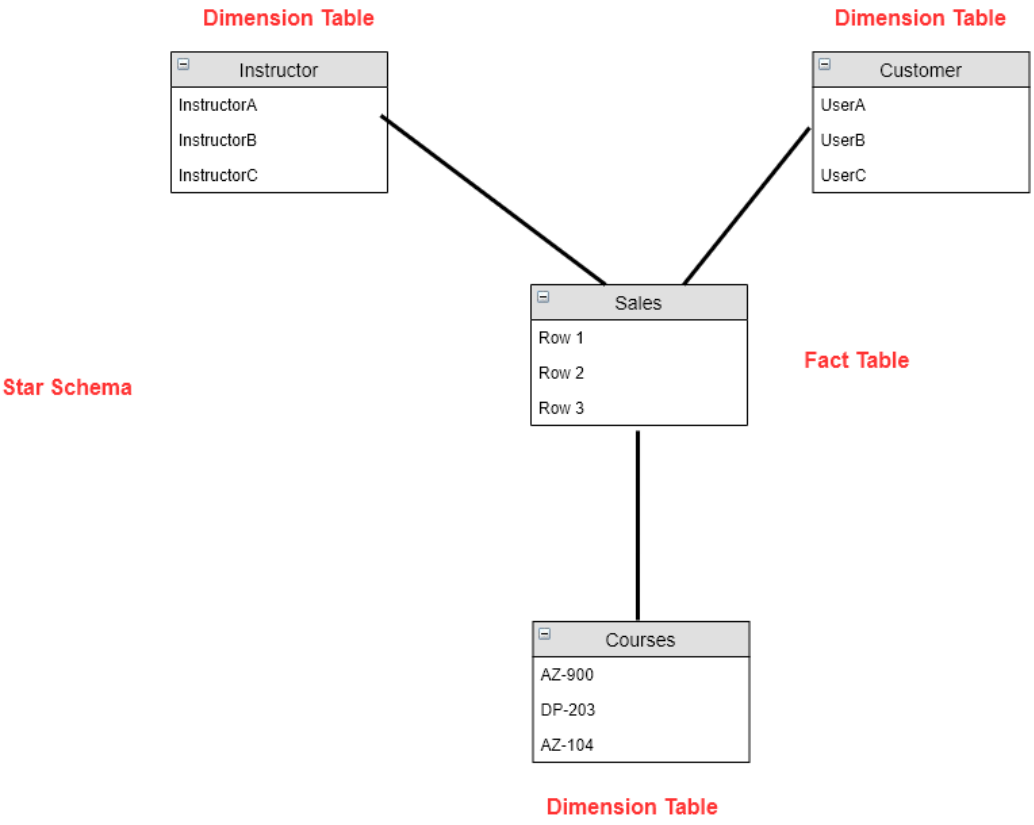
Sales
Row 1
Row 2
Row 3

Sales Fact table

Which courses sold the most in the year?

Which regions had the most customers?

Who are the most popular instructors?



Instructor
InstructorA
InstructorB
InstructorC

Customer
UserA
UserB
UserC

Courses
AZ-900
DP-200
DP-201

Date
20-10-2020
22-12-2020
23-12-2020

Normally the dimension tables will have a lot of attributes

The dimension tables will not contain that many rows

Normally rows are not added that frequently

InstrutorID
InstructorName
InstructorLocation
InstructorProfile

CustomerID
CustomerName
CustomerLocation
CustomerProfile

CourseID
CourseName
CoursePrice
LaunchDate

Who are the most popular instructors?

Which regions had the most customers?

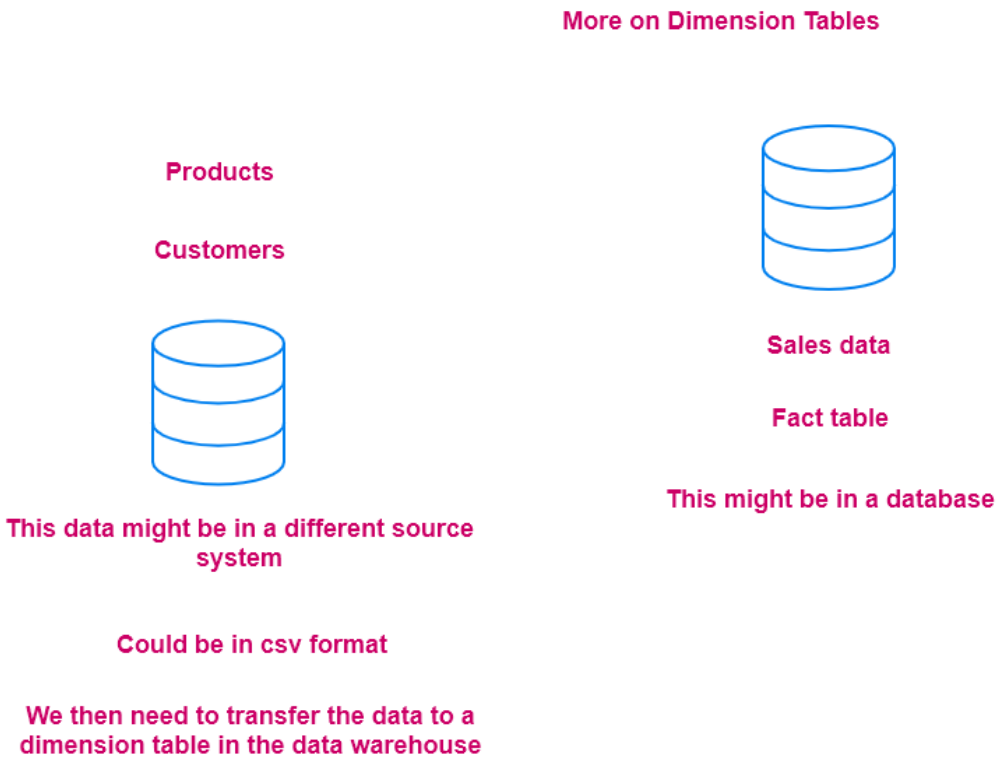
Which courses sold the most in the year?

Instructor Dimension Table
InstrutorID
InstructorName

Customer Dimension Table
CustomerID
CustomerLocation

Course Dimension Table
CourseID
CourseName

More on dimension tables



Source System A

ProductID	ProductName	ProductPrice
1	AZ-204	10.99
2	DP-203	11.99
3	AZ-104	12.99

Source System B

ProductID	ProductName	ProductPrice
1	BookA	10.99
2	BookB	11.99
3	BookC	12.99

One Product Dimension Table

TableKey	ProductID	ProductName	ProductPrice
1	1	AZ-204	10.99
2	2	DP-203	11.99
3	3	AZ-104	12.99
4	1	BookA	10.99
5	2	BookB	11.99
6	3	BookC	12.99

Here the Table Key is a surrogate key

Helps to uniquely identify each row in the table

You can use the Identity column feature in Azure Synapse for tables to generate the unique ID

Ideal practice

Don't have NULL values for properties in the dimension table. Won't give desired results when using reporting tools.

Try to replace the NULL value with some default value

Lab - Building a Fact Table

Fact tables

This is a usually large in size

These tables contain measurable facts

Sales made Courses purchased Number of orders

The fact table will contain the primary keys used in the dimension table

You can also create a surrogate key in the fact table to uniquely identify each row in the table

You can have NULL values for the facts

But don't have NULL values for the keys in the fact table that will be used for joins to the dimension tables

Understanding Azure Synapse Architecture

Dedicated SQL Pool - Data Warehouse

Hash-distributed tables | Replicated Tables | Round-robin distributed tables

Control Node

Compute Node

Compute Node

Compute Node

There are 60 distributions, but just showing 3 for simplicity

Distributions

The data is shared into distributions to optimize the peformance of work that can be done on the underlying data

Distributions are stored in Azure storage

Here your data and compute are seperate so that they can scale independently

The Control Node optimizes the query for parallel processing

The work is then passed to the Compute nodes

The Compute Nodes will do the work in parallel when it comes to the query

Understanding table types

Round-robin distributed tables

Id	Course	Category
1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

1	DP-203	Data

2	AZ-304	Architecture

3	AZ-303	Architecture

Hash-distributed tables

Id	Course	Category
1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

Here the data will be distributed based on the hash value calculated on a particular column

Lets say the hash is calculated based on the Category column

1	DP-203	Data

2	AZ-304	Architecture
3	AZ-303	Architecture

Replicated Tables

Here a full copy of the table is cached on each compute node

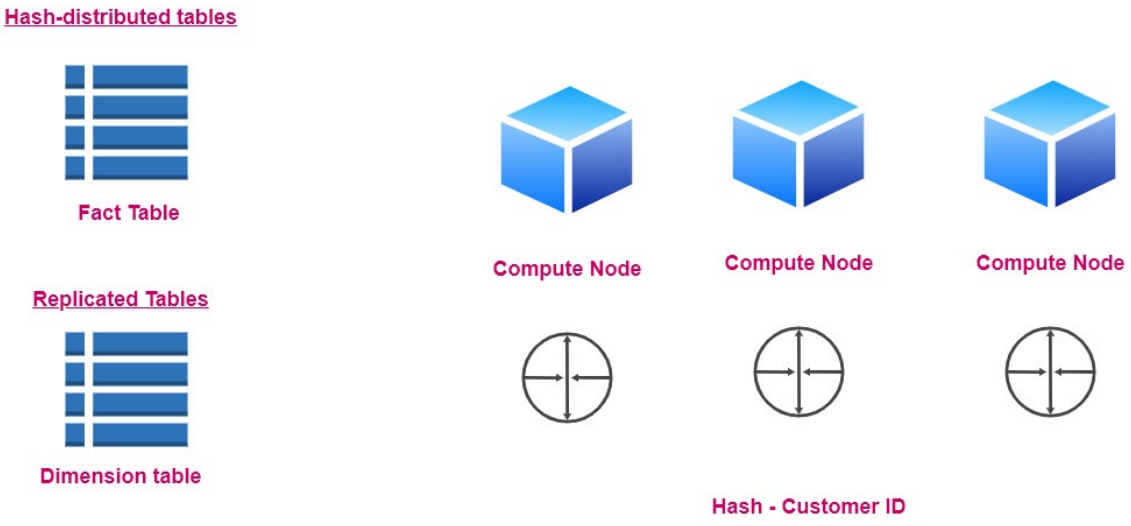
Id	Course	Category
1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

1	DP-203	Data
2	AZ-304	Architecture
3	AZ-303	Architecture

Designing your tables



Dedicated SQL Pool - Data Warehouse

Hash-distributed tables

Hash-distributed tables are good for large fact tables

This would be the ideal choice when you choose the right distribution column

For your distribution column ensure that it has many unique values so that data gets spread across more distributions

If not , then this may result in Data skew wherein data is not distributed across the distributions.

Don't use the date column for the hash-distributed column. This will make the data for the same date go to the same distribution

Round-robin distributed tables

This is not a good choice when you have frequent joins, because this would cause a lot of reshuffling of rows

This is a choice you can take if there is no good candidate column for the hash-distributed table

Also if you are planning on having a temporary staging table

Replicated Tables

These are good for your dimension tables in a star schema

Here since the tables are on each node, it would be faster when joins are performed with the fact and dimension tables

Ideal when the table size is less than 2 GB

May not be ideal if there are frequent insert, update or delete operations on the table

When you scale the SQL pool, then this table needs to be rebuilt on the nodes

Designing tables – Review

- **Hash-distributed table**
- A hash-distributed table takes the data and distributes the rows across the compute nodes. The distribution of data is done via a deterministic hash function that assigns each row to one distribution.
- Consider using hash-distributed tables if the table size is more than 2 GB on disk.
- **Choosing a distribution column**
- When choosing a column, you have to ensure to avoid data and processing skew.
- Data skew means the data is not distributed evenly across the distributions

Processing skew means that some distributions take longer than others when running parallel queries.

- **Hash-distributed table**
- Choose the column for distribution that has

- Many unique values
- Does not have NULLs or very few NULLs
- Is not the date column.
- Is used in JOIN, GROUP BY, HAVING clauses
- Is not used in the WHERE clauses.
- **Round-robin tables**
- This type of table distributes the rows evenly across all distributions.
- If a query needs data from different distributions, then data movement might be required to get the results of the query.
- If there are no joins performed on tables, then you can consider using this table type.
- Also, if there is no clear candidate column for hash distributing the table.
- Normally temporary staging tables can use this table type.
- **Replicated tables**
- Here each compute node has the full copy of the table.
- Consider the use of replicated tables when the table size is less than 2 GB compressed.
- Its good to consider this table type for your dimension tables.
- Don't consider this table type if the table has frequent insert, update, and delete operations. This could require a rebuild of the replicated table.
- Use replicated tables for queries with simple query predicates, such as equality or inequality.
- Use distributed tables for queries with complex query predicates, such as LIKE or NOT LIKE.

Lab - Windowing Functions

- A Windowing function allows one to apply a mathematical equation on a set of data that is defined within a window.

- With the function, you can split the rows of data into different sets and then apply an aggregate to the data in each set.
- When using windowing functions with SQL Pools, you will use the OVER clause.
- This clause determines the partitioning and ordering of a rowset before the associated window function is applied.

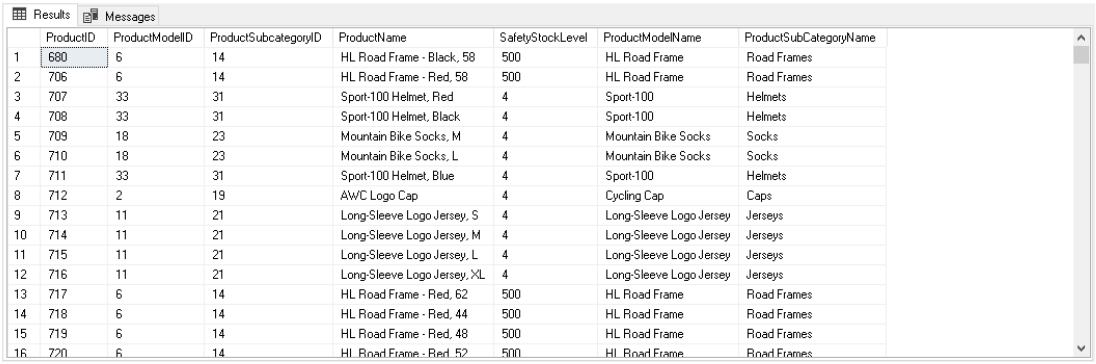
Then you can apply the aggregate function accordingly

Lab - Surrogate keys for dimension tables

Surrogate Keys

Here the ProductID is referred to as the Alternate Key or the Business Key

This refers to the Primary Key in the source system



Results | Messages

	ProductID	ProductModelID	ProductSubcategoryID	ProductName	SafetyStockLevel	ProductModelName	ProductSubCategoryName
1	680	6	14	HL Road Frame - Black, 58	500	HL Road Frame	Road Frames
2	706	6	14	HL Road Frame - Red, 58	500	HL Road Frame	Road Frames
3	707	33	31	Sport-100 Helmet, Red	4	Sport-100	Helmets
4	708	33	31	Sport-100 Helmet, Black	4	Sport-100	Helmets
5	709	18	23	Mountain Bike Socks, M	4	Mountain Bike Socks	Socks
6	710	18	23	Mountain Bike Socks, L	4	Mountain Bike Socks	Socks
7	711	33	31	Sport-100 Helmet, Blue	4	Sport-100	Helmets
8	712	2	19	AWC Logo Cap	4	Cycling Cap	Caps
9	713	11	21	Long-Sleeve Logo Jersey, S	4	Long-Sleeve Logo Jersey	Jerseys
10	714	11	21	Long-Sleeve Logo Jersey, M	4	Long-Sleeve Logo Jersey	Jerseys
11	715	11	21	Long-Sleeve Logo Jersey, L	4	Long-Sleeve Logo Jersey	Jerseys
12	716	11	21	Long-Sleeve Logo Jersey, XL	4	Long-Sleeve Logo Jersey	Jerseys
13	717	6	14	HL Road Frame - Red, 62	500	HL Road Frame	Road Frames
14	718	6	14	HL Road Frame - Red, 44	500	HL Road Frame	Road Frames
15	719	6	14	HL Road Frame - Red, 48	500	HL Road Frame	Road Frames
16	720	6	14	HL Road Frame - Red, 52	500	HL Road Frame	Road Frames

In Dimension tables, you will also want to have a surrogate key

The Surrogate key is also sometimes referred to as the non-business key

This can be simple incrementing integer values

In the SQL pool tables, you can use the IDENTITY column feature

Slowly Changing dimensions

Slowly changing dimensions

If the ProductName changes in the source table , then this change needs to be reflected in the Dimension table

Type 1

Here you just update the changes as they are

Results | Messages

	ProductID	ProductModelID	ProductSubcategoryID	ProductName	SafetyStockLevel	ProductModelName	ProductSubCategoryName
1	680	6	14	HL Road Frame - Black, 58	500	HL Road Frame	Road Frames
2	706	6	14	HL Road Frame - Red, 58	500	HL Road Frame	Road Frames
3	707	33	31	Sport-100 Helmet, Red	4	Sport-100	Helmets
4	708	33	31	Sport-100 Helmet, Black	4	Sport-100	Helmets
5	709	18	23	Mountain Bike Socks, M	4	Mountain Bike Socks	Socks
6	710	18	23	Mountain Bike Socks, L	4	Mountain Bike Socks	Socks
7	711	33	31	Sport-100 Helmet, Blue	4	Sport-100	Helmets
8	712	2	19	AWC Logo Cap	4	Cycling Cap	Caps
9	713	11	21	Long-Sleeve Logo Jersey, S	4	Long-Sleeve Logo Jersey	Jerseys
10	714	11	21	Long-Sleeve Logo Jersey, M	4	Long-Sleeve Logo Jersey	Jerseys
11	715	11	21	Long-Sleeve Logo Jersey, L	4	Long-Sleeve Logo Jersey	Jerseys
12	716	11	21	Long-Sleeve Logo Jersey, XL	4	Long-Sleeve Logo Jersey	Jerseys
13	717	6	14	HL Road Frame - Red, 62	500	HL Road Frame	Road Frames
14	718	6	14	HL Road Frame - Red, 44	500	HL Road Frame	Road Frames
15	719	6	14	HL Road Frame - Red, 48	500	HL Road Frame	Road Frames

Type 2

Here you keep both the OLD and NEW values in the Dimension table

ProductSK	ProductID	ProductName	StartDate	EndDate	IsCurrent
1	1	ProductA	2021-03-20	2021-04-20	False
2	1	ProductAA	2021-04-21	9999-12-31	True

Type 3

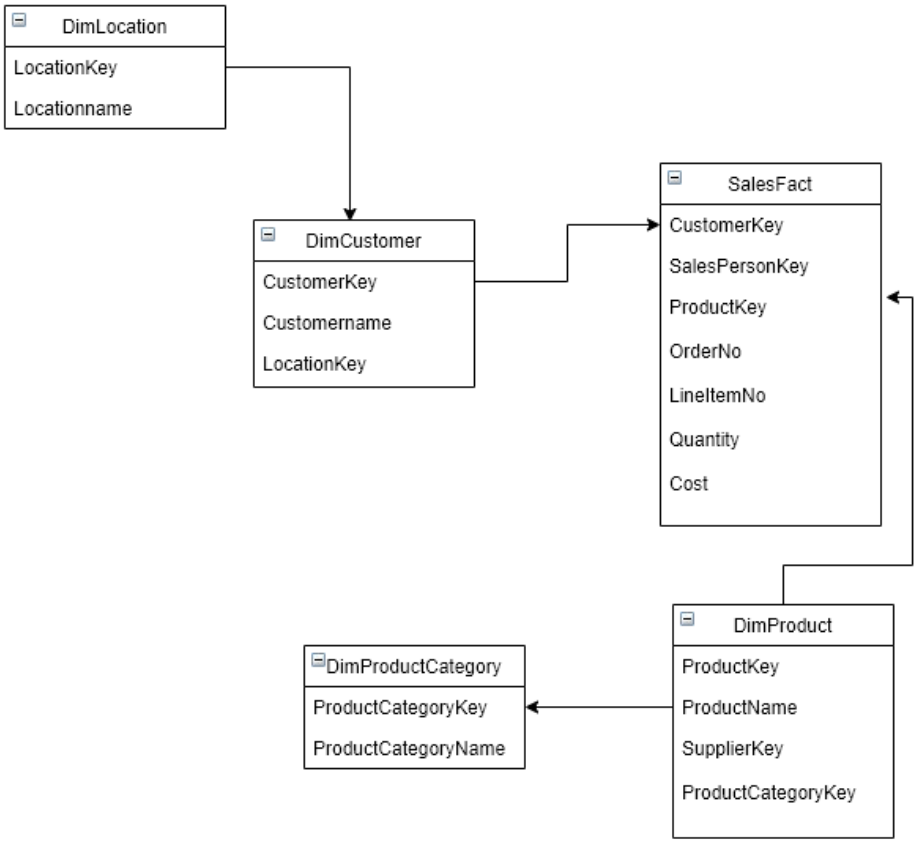
Here instead of having multiple rows to signify changes, we now have additional columns to signify the changes

ProductSK	ProductID	Original Name	Changed Name	EffectiveDate
1	1	ProductA	ProductAA	2021-04-20

Snowflake schema

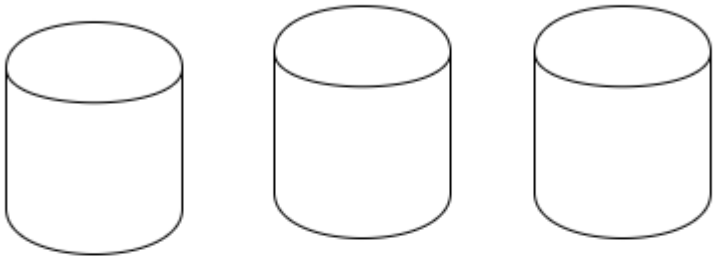
Snowflake schema

Set of normalized tables when it comes to the dimension tables



Partitions in Azure Synapse

Table Partitions



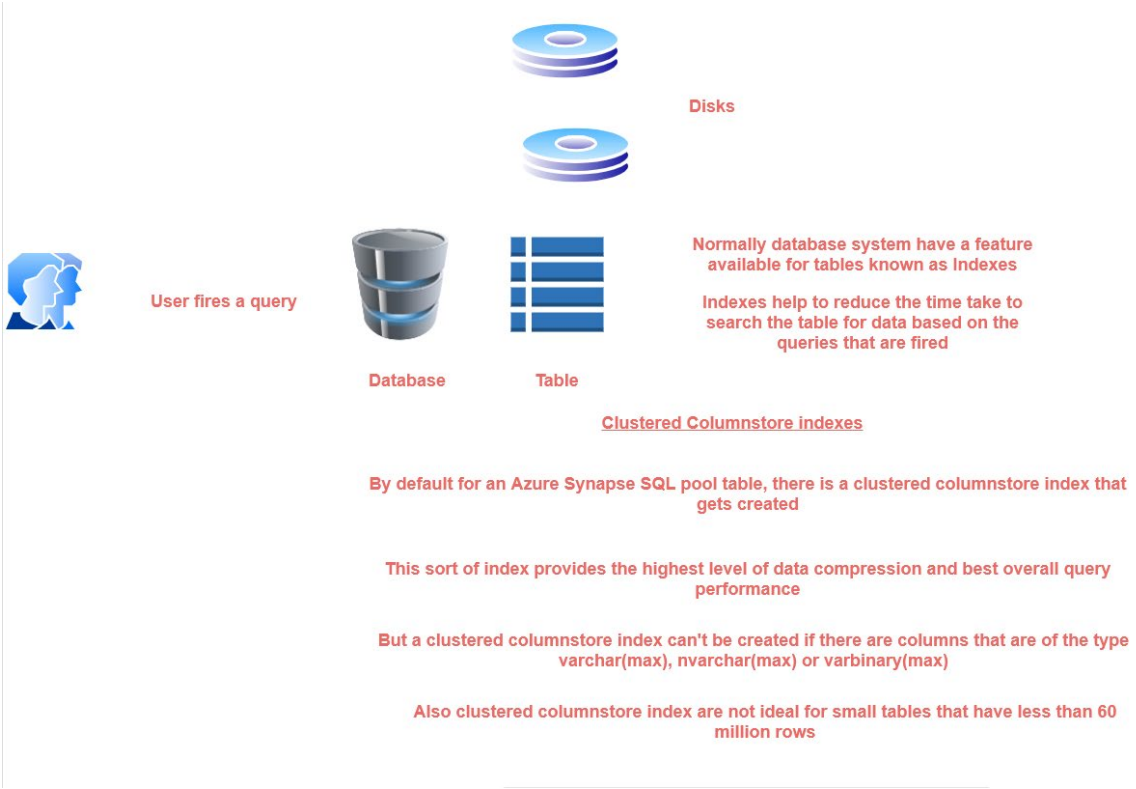
This helps to divide data into
smaller groups of data

Normally data is partitioned by dates

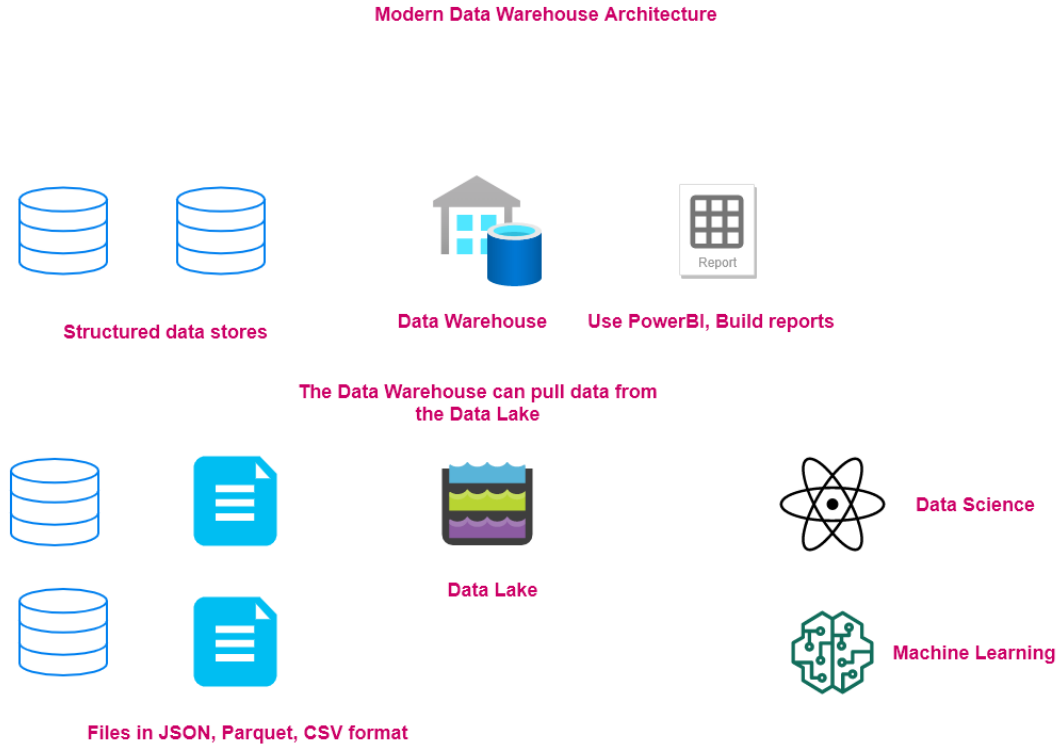
Partitioning also helps in filtering data when using the WHERE clause in your queries

Here the engine can then just process the data in the partitions based on the condition mentioned in the WHERE clause

Indexes



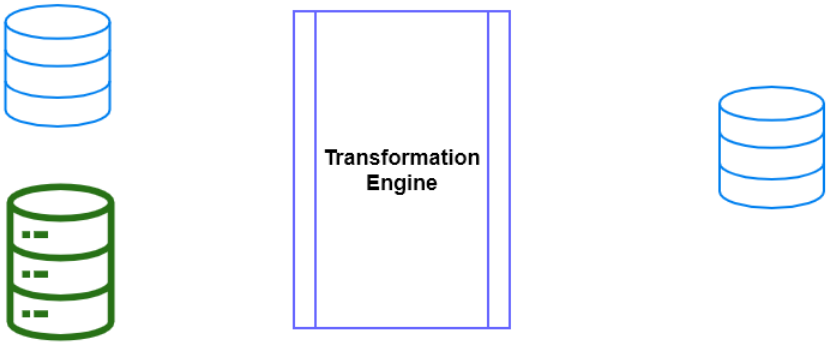
Quick Note - Modern Data Warehouse Architecture



Design and Develop Data Processing - Azure Data Factory

Extract, Transform and Load

Extract, transform and load process (ETL)



Extract data from various data sources

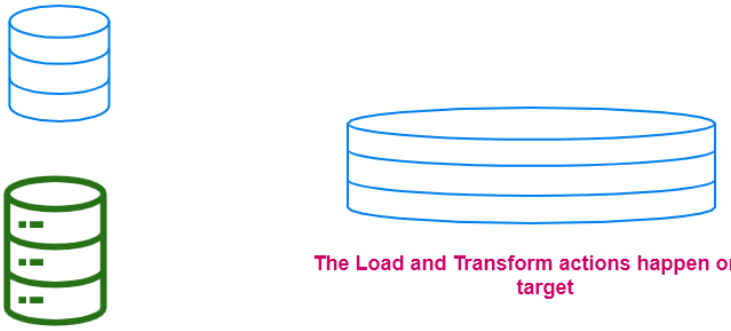
Load the data into the target

Here you need to have a seperate transformation engine to transform the data

Various operations performed by the engine - filtering, sorting, aggregating, joining data, cleaning data

Tools - Azure Data Factory, SQL Server Integration Services

Extract, load and transform process (ELT)



The Load and Transform actions happen on the target

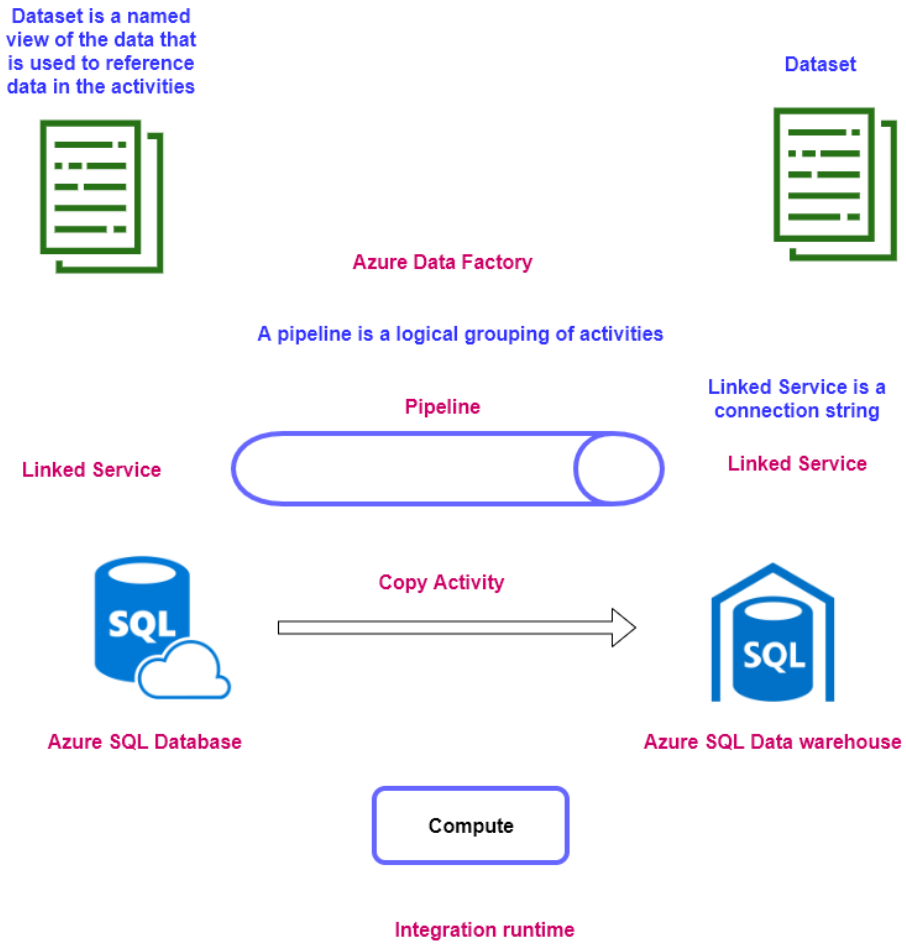
Extract data from various data sources

Here you dont need a seperate transformation engine

Here the target system must be powerful enough to do the transformation

Tools - Azure Synapse, Azure Data Factory

What is Azure Data Factory



- This is a cloud based ETL tool (Extract, Transform, Load)
- Here you can create data-driven workflows.
- These workflows help to orchestrate data movement.
- It an also help to transform the data.

- **<u>The Azure Data Factory process</u>**
- The first step is to connect to the required data sources.
- The next step is to ingest the data from the source.
- You can then transform the data in the pipeline if required.
- You can then publish the data onto a destination – Azure Data Warehouse, Azure SQL Database , Azure Cosmos DB.

You can also monitor the pipeline as it is running.

- **The Azure Data Factory components**
- **Linked Service** – This enables you to ingest data from a data source. The Linked Service can create the required compute resources to take the data from the data source.
- **Datasets** – This represents the data structure within the data store that is being referenced by the Linked Service object.
- **Activity** – This contains the actual transformation logic. You can also have simple Copy Activities to copy the data from the source to the destination.

Mapping Data Flow

- This helps to visualize the data transformations in Azure Data Factory.
- Here you can write the required transformation logic without actually writing any code.
- The data flows are run on Apache Spark clusters.
- Here Azure Data Factory will handle the transformations in the data flow.
- **Debug mode** – You also have a Debug mode in place. Here you can actually see the results of each transformation.
- In the debug mode session, the data flow is run interactively on a Spark cluster.

In the debug mode you will be charged on an hourly basis that the cluster is active.

Self-Hosted Integration Runtime

Self-Hosted Integration runtime

Azure Data Lake

Azure Data Factory

Azure Synapse

Your data might be hosted on a virtual machine

The virtual machine could be hosted in your on-premises infrastructure

It could have data files or a SQL database

To register the server , you need to install the self-hosted integration runtime

Azure Data Factory

Azure Synapse

Lab - Azure DevOps - Git configuration

Developers

Code

Git

Version control

Cloud versions

Github

Azure Repos

Pipelines can also be versioned controlled

Lab - Azure DevOps - Release configuration

DevOps - Integrating Azure Data Factory

We have already seen how to version control DevOps pipelines in Azure repos

What is the normal cycle for Continuous Integration/Delivery

Here your developers/data engineers have complete access to work with Azure Data Factory

Developement resource for Azure Data Factory

Feature branch

All developers/engineers will make changes to the branch

Main branch

This is also known as the collaboration branch

One the changes are complete , a pull request is created to merge the changes into the main branch

This allows the changes to be reviewed by peers

Then the changes are published to Azure Data Factory

And finally the pipeline can be promoted to another instance of Azure Data Factory

Azure Data Factory Development

Azure Data Factory Staging

Azure Data Factory Production

Design and Develop Data Processing - Azure Event Hubs and Stream Analytics

Batch and Real-Time Processing

Batch Processing

Here you take large amounts of data from the source, transform it and load into a destination data store

You might take a large set semi-structured files and then process then into structured files that can be used for future Analysis needs

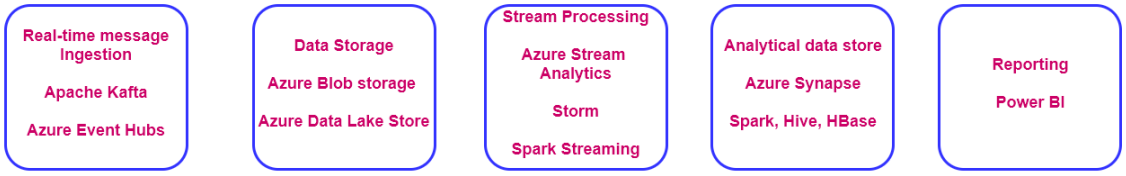
Example - Web Server logs that are copied to Azure Data Lake Gen2 storage accounts over the day. And then a batch process kicks in the night that would process the data and send it to an Analytical store for daily reporting purposes.

Real time Processing

Here streams of data are captured in real-time and processed with minimal latency to generate real-time reports

Here processing of data needs to be done as fast as possible so that it does not block the incoming stream of data

Also platforms should be available to ingest large amounts of data at a fast rate



What are Azure Event Hubs

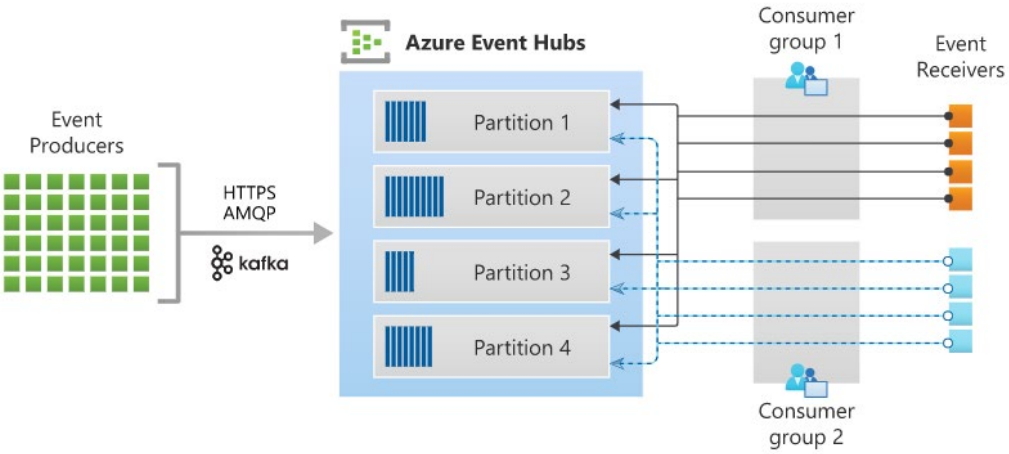
What are Azure Event Hubs

This is a big data streaming platform

This service can receive and process millions of events per second

You can stream log data , telemetry data, any sort of events to Azure Event Hubs

Event Hubs Architecture



The different components

Event producers - This is an entity that sends data to an event hub. The events can be published using the protocols - HTTPS, AMQP, Apache Kafka

Partitions - The data is split across partitions. This allows for better throughput of your data onto Azure Event Hubs

Consumer groups - This is a view (state,position or offset) of an entire event hub

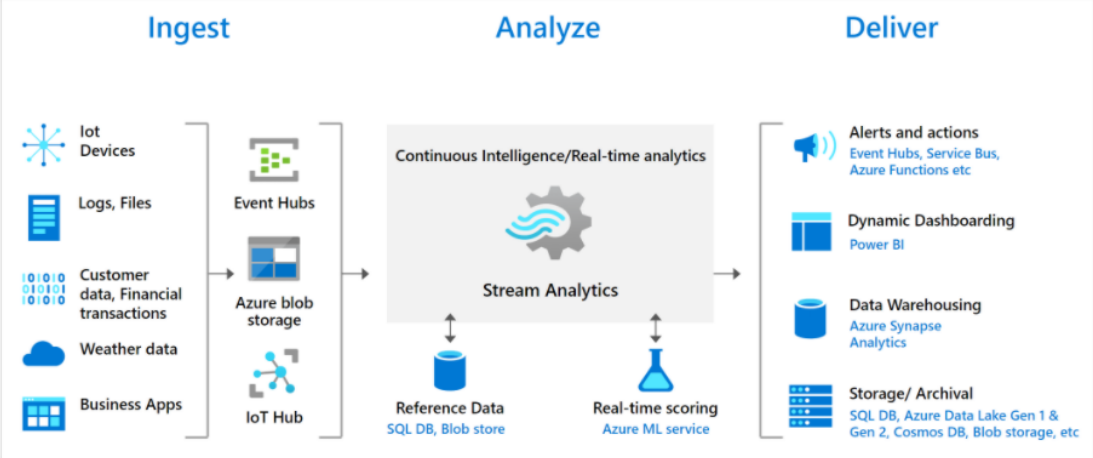
Throughput - This controls the throughput capacity of Event Hubs

Event Receivers - This is an entity that reads event data

What is Azure Stream Analytics

Azure Stream Analytics

This is a real-time analytics and event-processing service



https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

Lab - Reference data

Using Reference data

Azure Blob

Azure Stream Analytics

Azure SQL database

Sometimes your query might want to reference static data that rarely changes

You can perform a JOIN on your streaming data to your reference data

Your reference data can be a blob in a storage accout or data in an Azure SQL database

Lab - Azure Event Hubs - Capture Feature

Azure Event Hubs Capture

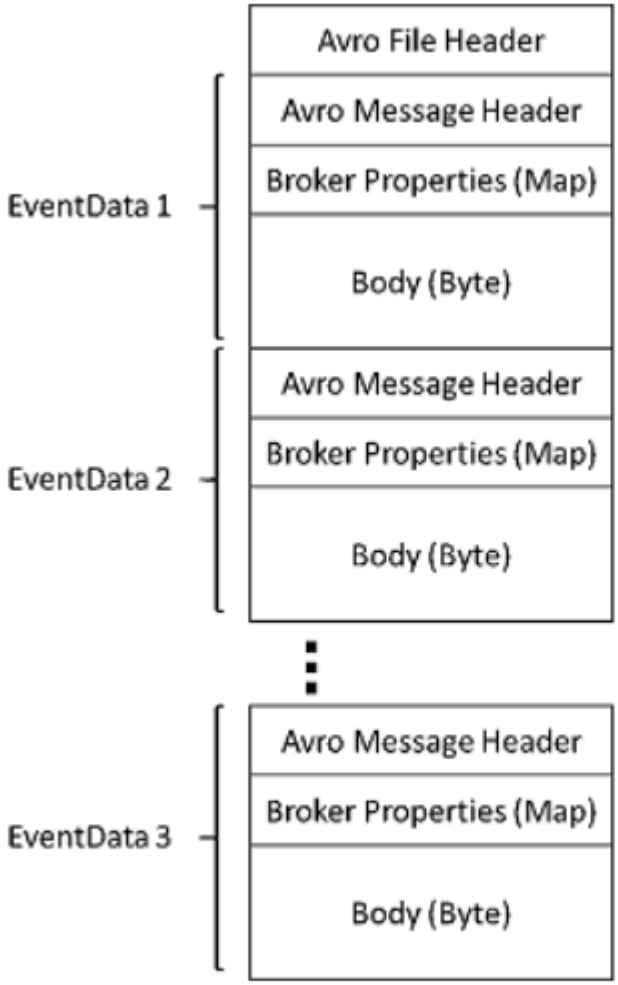
Azure Event Hubs

Azure Stream Analytics

Azure Data Lake

Azure Data Factory

With Azure Event Hubs Capture, you can stream events onto Azure Blob storage or Azure Data Lake Storage Accounts



https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-capture-overview

Design and Develop Data Processing - Scala, Notebooks and Spark

Spark Pool

Spark Dataset

- This is a strongly typed collection of domain-specific objects.
- This data can then be transformed in parallel.
- Normally you will perform either transformations or actions on a dataset.
- The transformation will produce a new dataset.

› The action will trigger a computation and produce the required result.

The benefit of having a Dataset is that you can use powerful transformations on the underlying data.

Spark DataFrame

› The DataFrame is nothing but a Dataset that is organized into named columns.

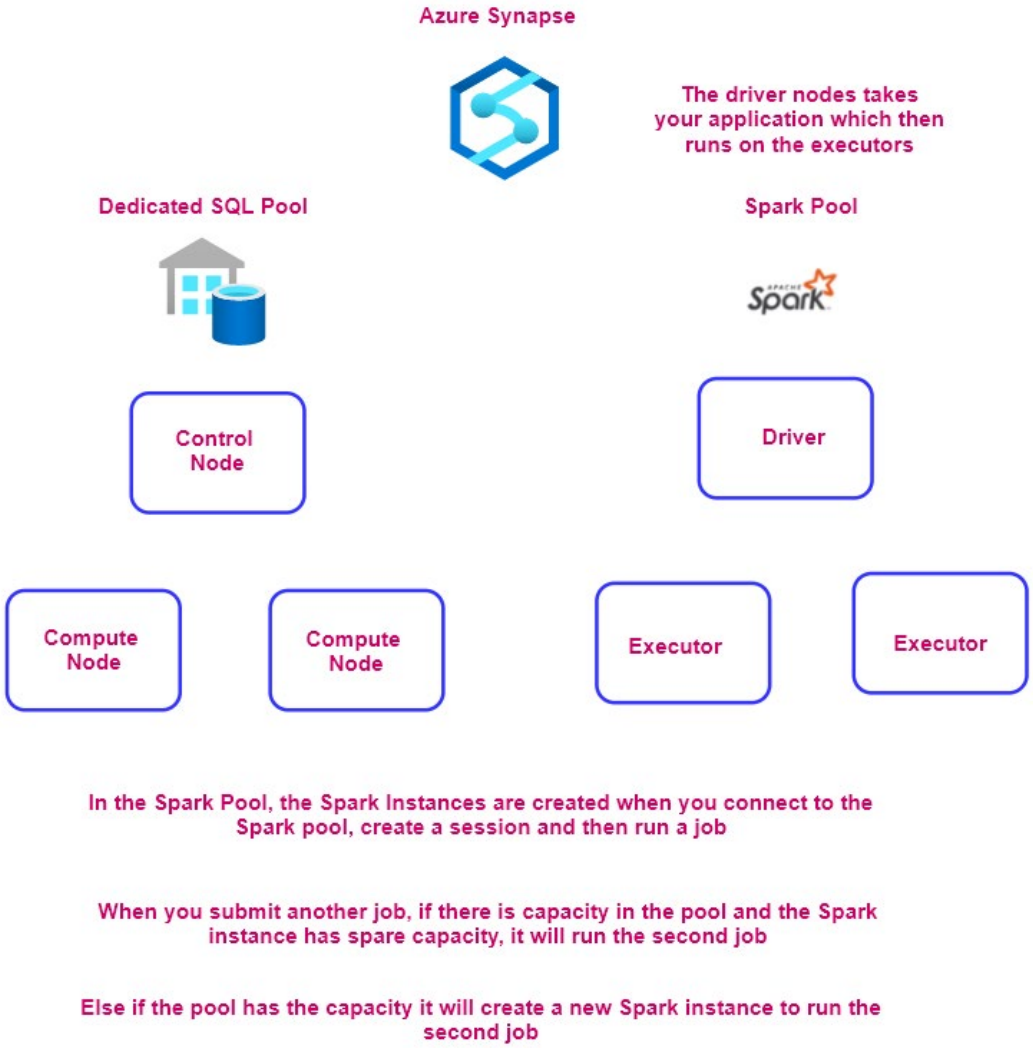
› Its like a table in a relational database.

› You can construct DataFrames from external files.

› When it comes to Datasets, the API for working with Datasets is only available for Scala and Java.

› For DataFrames, the API is available in Scala, Java, Python and R.

Spark Pool - Combined Power



Design and Develop Data Processing - Azure Databricks

What is Azure Databricks

Databricks

This is a company that was founded by the original creators of Apache Spark

Databricks makes use of Apache Spark to provide a Unified Analytics platform

You would need to provision the machines

Install Spark and the necessary libraries

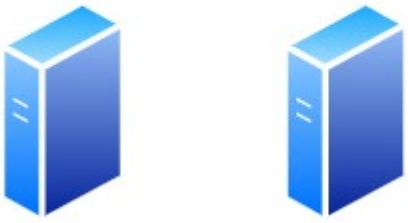
Maintain the scaling and availability of the machines

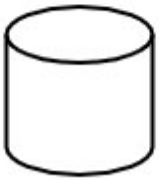
With Databricks , the entire environment can be provisioned with just a few clicks

What Databricks does for you

1. It will create the underlying compute infrastructure for you

2. It has its own underlying file system which is an abstraction of the underlying storage layer





3. It will install Spark for you. It also has the capabilities to install other libraries and frameworks - Machine Learning libraries

4. It provides a workspace for you

In the workspace you can work with Notebooks

Users can collaborate on the Notebooks

Create visualizations on the Notebook

Azure Databricks

Completely managed enviroment for you

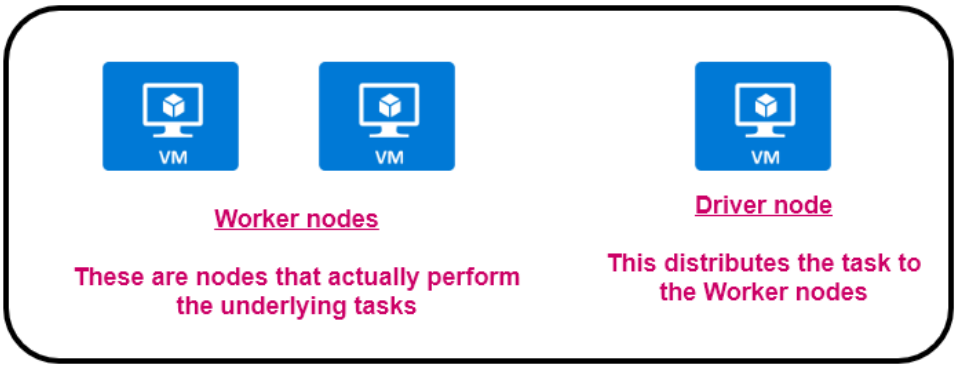
Makes use of the underlying compute infrastructure and virtual networks

Make use of Azure security - Azure Active Directory and Role-based access control

Clusters in Azure Databricks

The cluster will have the Spark engine and other components installed

Cluster



Worker nodes

These are nodes that actually perform the underlying tasks

Driver node

This distributes the task to the Worker nodes

The workloads can run on the cluster via a set of commands

Interactive Cluster

Here you can analyze data with the help of interactive notebooks

Multiple users can use a cluster and collaborate

Job cluster

Here the work runs as a job on the cluster

When a job has a to run, Azure Databricks will start the cluster

When the job is complete the cluster will be terminated

Two types of Interactive Cluster

Standard cluster	High concurrency cluster
This is recommended if you are a single user	This is recommended for multiple users
Here there is no fault isolation. If multiple users are using a cluster and one user has a fault, it can impact the workloads of other users	Here you have fault isolation
Here the resources of the cluster might get allocated to a single workload	Here the resources of the cluster are effectively shared across the different user workloads
Has support for Python, R , Scala and SQL	Has support for Python, R and SQL
	Table Access Control - Here you can grant and revoke access to data from Python and SQL

Databricks File System

Your workspace has a Databricks File System

This is an abstraction layer on top of scalable object storage

This allows you to interact with object storage using directory and file semantics

These files persists even if the cluster is terminated

The default storage location in DBFS is called DBFS root

There are some predefined DBFS root locations

/FileStore	Imported data files, generated plots, uploaded libraries
/databricks-datasets	Sample public datasets
/user/hive/warehouse	Data and metadata for non-external Hive tables

Autoscaling a cluster

- When creating an Azure Databricks cluster, you can specify a minimum and maximum number of workers for the cluster.
- Databricks will then choose the ideal number of workers to run the job.
- If a certain phase of your job requires more compute power, the workers will be assigned accordingly.
- There are two types of autoscaling
- Standard autoscaling
- Here the cluster starts with 8 nodes

- Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.

Scales down exponentially, starting with 1 node

- **Optimized autoscaling**
- This is only available for Azure Databricks Premium Plan.
- Can scale down even if the cluster is not idle by looking at shuffle file state.
- Scales down based on a percentage of current nodes.
- On job clusters, scales down if the cluster is underutilized over the last 40 seconds.
- On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

Lab - Azure Databricks Table

- In Azure Databricks, you can also create a database and tables.
- The table is a collection of structured data.
- You can then perform operations on the data that are supported by Apache Spark on DataFrames on Azure Databricks tables.
- There are two types of tables – global and local tables. A global table is available across all clusters.
- A global table is registered in the Azure Databricks Hive metastore or an external metastore.
- The local table is not accessible from other clusters and is not registered in the Hive metastore.

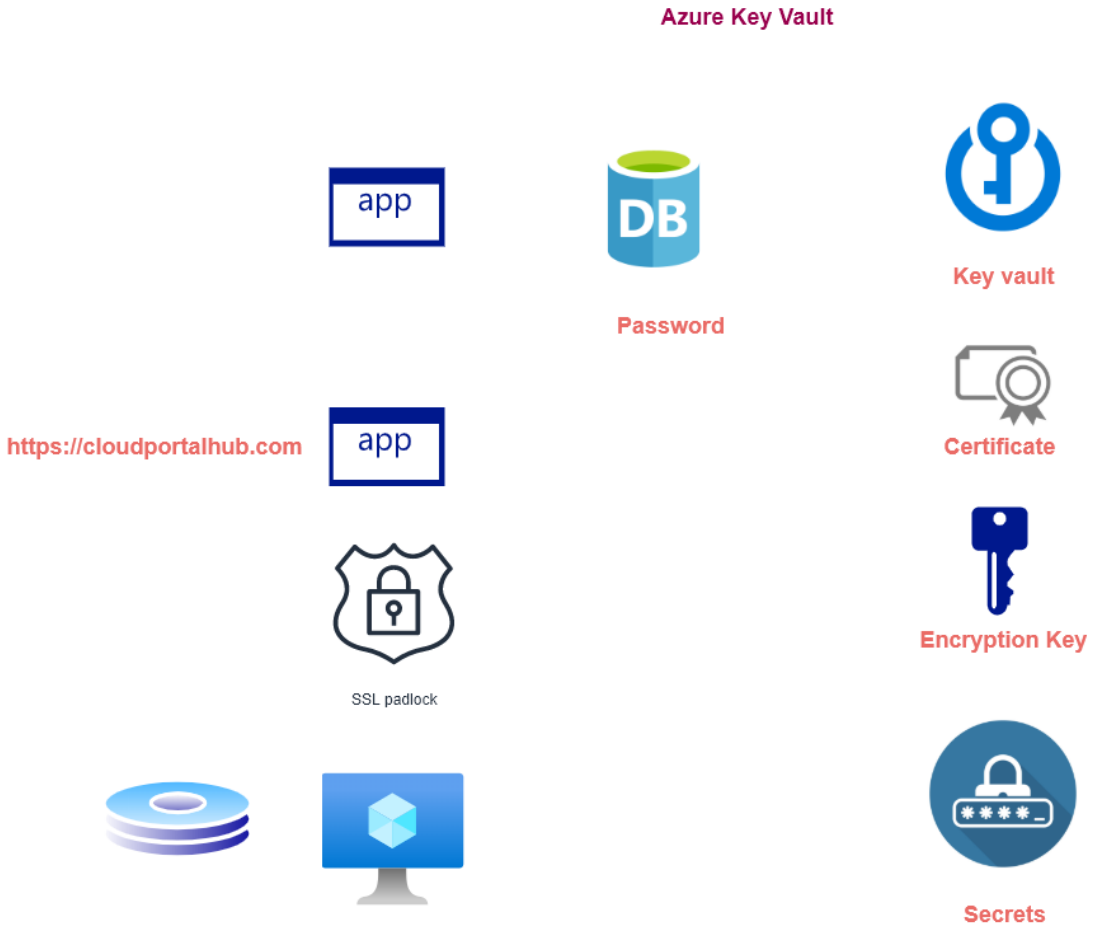
Delta Lake Introduction

- **ACID transactions on Spark** - Serializable isolation levels ensure that readers never see inconsistent data.
- **Scalable metadata handling** - Leverages Spark distributed processing power to handle all the metadata for petabyte-scale tables with billions of files at ease.

- **Streaming and batch unification** - A table in Delta Lake is a batch table as well as a streaming source and sink. Streaming data ingest, batch historic backfill, interactive queries all just work out of the box.
- **Schema enforcement** - Automatically handles schema variations to prevent insertion of bad records during ingestion.
- **Time travel** - Data versioning enables rollbacks, full historical audit trails, and reproducible machine learning experiments.
- **Upserts and deletes** - Supports merge, update and delete operations to enable complex use cases like change-data-capture, slowly-changing-dimension (SCD) operations, streaming upserts, and so on.

Design and Implement Data Security

What is the Azure Key Vault service



Azure Data Factory – Encryption

- Azure Data Factory already encrypts data at rest which also includes entity definitions and any data that is cached.
- The encryption is carried out with Microsoft-managed keys.
- But you can also define your own keys using the Azure Key vault service.
- For the key vault, you have to ensure that Soft delete is enabled and the setting of Do Not Purge is also enabled.
- Also grant Azure Data Factory the key permissions of 'Get', 'Unwrap Key' and 'Wrap Key'

Lab - Azure Synapse - Data Masking

- Here the data in the table can be limited in its exposure to non-privileged users.
- You can create a rule that can mask the data.
- Based on the rule you can decide on the amount of data to expose to the user.
- There are different masking rules.
- **Credit Card masking rule** – This is used to mask the column that contain credit card details. Here only the last four digits of the field are exposed.
- **Email** – Here first letter of the email address is exposed. And the domain name of the email address is replaced with XXX.com.
- **Custom text**- Here you decide which characters to expose for a field.
- **Random number**- Here you can generate a random number for the field.

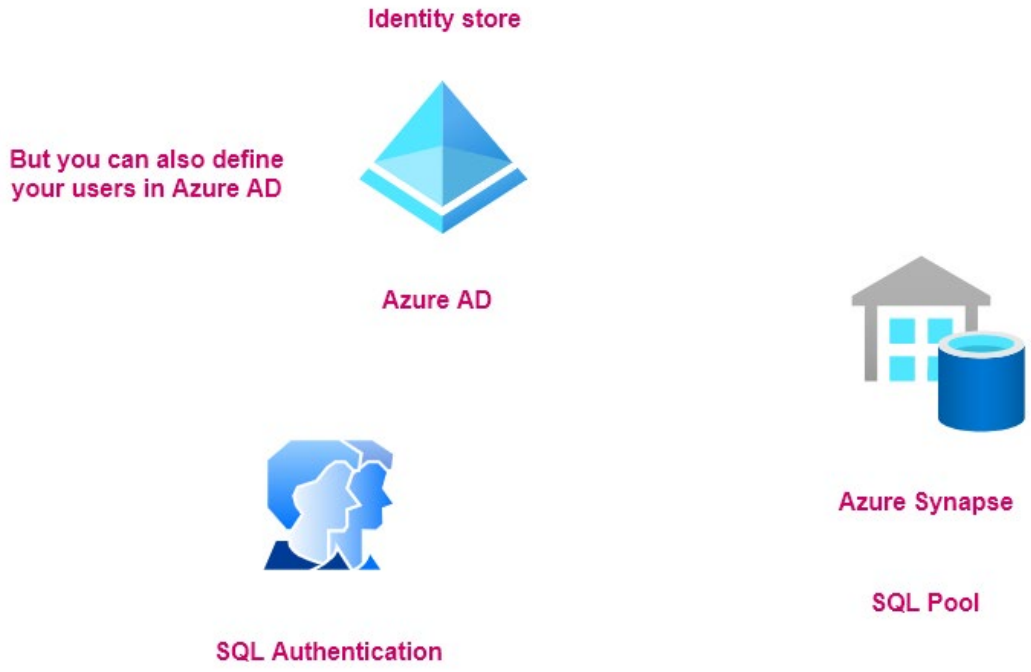
Lab - Azure Synapse - Auditing

- You can enable auditing for an Azure SQL Pool in Azure Synapse Analytics.
- This feature can be used to track database events and write them to an audit log.
- The logs can be stored in an Azure storage account, a Log Analytics workspace and Azure Event Hubs.
- This helps in regulatory compliance. It helps to gain insights on any anomalies when it comes to database activities.
- Auditing can be enabled at the data warehouse level or server level.
- If it is applied at the server level, then it will be applied to all of the data warehouses that reside on the server.

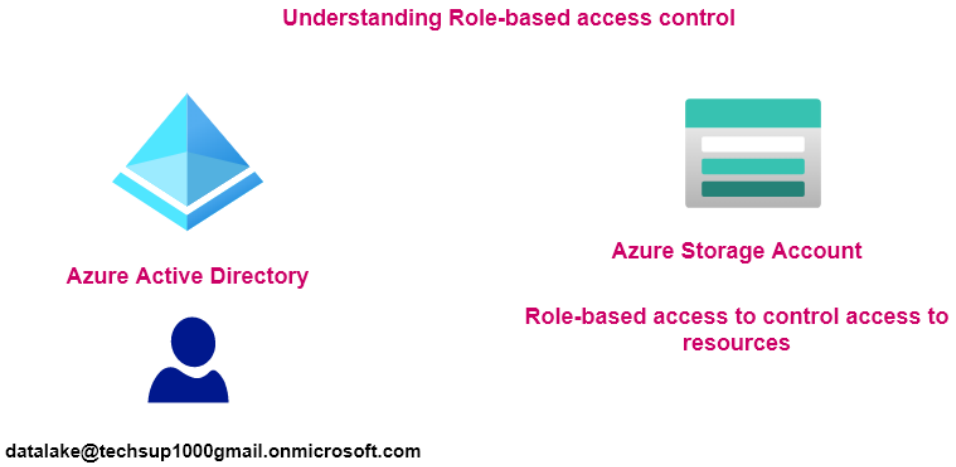
Azure Synapse - Data Discovery and Classification

- This feature provides capabilities for discovering, classifying, labelling, and reporting the sensitive data in your databases.
- The data discovery feature can scan the database and identify columns that contains sensitive data. You can then view and apply the recommendations accordingly.
- You can then apply sensitivity labels to the column. This helps to define the sensitivity level of the data stored in the column.

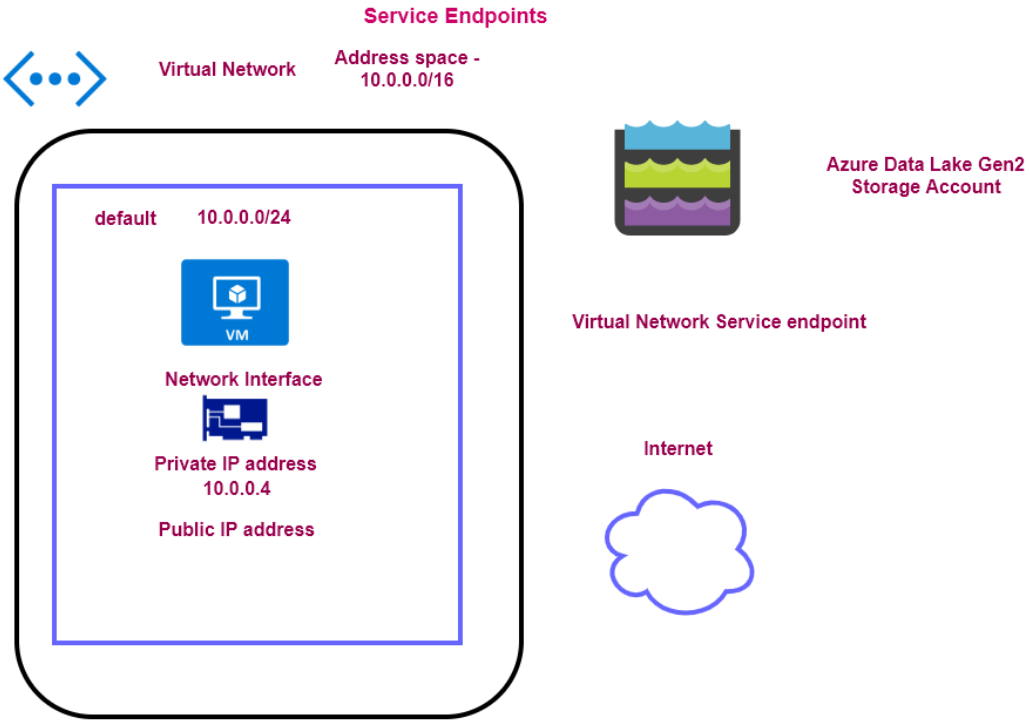
Azure Synapse - Azure AD Authentication



Lab - Azure Data Lake - Role Based Access Control



Lab - Azure Data Lake - Virtual Network Service Endpoint



Monitor and optimize data storage and data processing

Best practices for structing files in your data lake

Best practices for structuring files

1. Well-engineered data lake structure

Normally when designing a data-lake you might create multiple zones

The zones can map to seperate containers

RAW Zone

This contains files in its original format Avro, Parquet, JSON etc

Filtered Zone

Here basic filtering has been carried out. Un-necessary columns removed

Curated Zone

This is the data on which you might perform analytics on

Hierarchy used for the storage of files

\Department\RAW\DataSource\YYYY\MM\DD\File.json

2. Compress files

Use compressed file formats like Parquet.

Less time is spent on data transfer

The MPP architecture of the data warehouse can be used for decompression

3. Use multiple source files

Split your source files into different parts

If you have multiple compute nodes, each node can process one file

Azure Synapse - Workload Management

SQL Pool

You can have different types of workloads running on the SQL Pool

One set of users loading data into the SQL Pool

One set of users performing analysis of the data

Most of the time you need to manage the workloads so that resources are managed accordingly.

First you can create a Workload Group

This forms the basis of workload management

Here you define the boundaries of resources for the workload group

```
CREATE WORKLOAD GROUP DataLoads
WITH (
    MIN_PERCENTAGE_RESOURCE = 100
    ,CAP_PERCENTAGE_RESOURCE = 100
    ,REQUEST_MIN_RESOURCE_GRANT_PERCENT = 100
    );
```

And you define which users or roles are part of the group via a classifier

```
CREATE WORKLOAD CLASSIFIER [ELTLogin]
WITH (
        WORKLOAD_GROUP = 'DataLoads'
    ,MEMBERNAME = 'user_load'
);
```

Azure Synapse - Retention points

- Regular backups are for your dedicated SQL pool. These are snapshots of the data warehouse that are taken throughout the day.
- These restore points are then available for 7 days.
- You can restore your data warehouse in the primary region from any one of the snapshots taken in the past seven days.

You can also define your own user-defined snapshots.

Azure Key Vault - High Availability

.Net client library

Azure key vault high availability

East US

West US

The contents of the key vault are replicated within the region and to a secondary region as defined by Azure paired regions

In the event the primary region goes down, the requests for the key vault will failover to the secondary region. It takes a few minutes for the failover to take place.

The vault will be in read-only mode during the failover

Azure Stream Analytics – Metrics

- Backlogged Input Events - Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. For this you can decide to scale up the number of streaming units.
- Data Conversion Errors - Number of output events that could not be converted to the expected output schema.
- Early Input Events - Events whose application timestamp is earlier than their arrival time by more than 5 minutes.
- Late Input Events - Events that arrived later than the configured late arrival tolerance window.

- Out-of-Order Events - Number of events received out of order that were either dropped or given an adjusted timestamp, based on the Event Ordering Policy.
- **Watermark Delay** - The maximum watermark delay across all partitions of all outputs in the job.
- If the value of the Watermark Delay is greater than 0, it could be due to many reasons.
- Inherent processing delay of the streaming pipeline.
- Clock skew of the processing node generating the metric.

There are not enough processing resources in your Stream Analytics job.

Azure Stream Analytics - Streaming Units

Azure Stream Analytics job

When you create a Stream Analytics job, you assign a number of Streaming Units

The streaming units determines the computing resources that are allocated to execute the job

To ensure low latency when it comes to stream processing, all of the jobs are performed in memory

If the Streaming Units utilization reaches 100% , then your jobs will start failing

Hence always ensure to monitor the streaming units being consumed for a job

Standard streaming unit

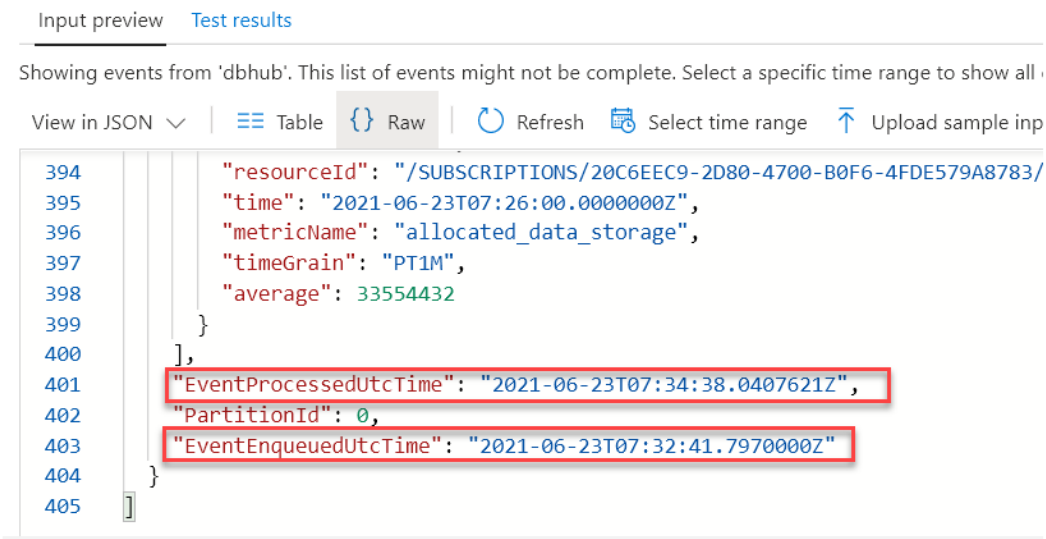
	Standard	Dedicated
Resource Type	Stream Analytics Job	Stream Analytics Cluster
Streaming unit	$0.11/hour with a 1 SU minimum	$0.11/hour with a 36 SU minimum*
Virtual Network support	No	Yes
C# User-defined functions	Limited to West Central US, North Europe, East US, West US, East US 2 and West Europe	All regions
Custom deserializers	Limited to West Central US, North Europe, East US, West US, East US 2 and West Europe	All regions

https://azure.microsoft.com/en-us/pricing/details/stream-analytics/

Azure Stream Analytics - The importance of time

```
SELECT time,COUNT(*) AS [Count]
INTO
    summary
FROM
    staginglogs
GROUP BY time,TumblingWindow(second,10)
```

Here we are grouping by the time attribute that is present in the event



Apart from that we also have the properties of EventProcessedUtcTime and EventEnqueuedUtcTime

Azure SQL database

Application time - This is when the application generated the event

Azure Event Hub

Arrival Time - This is when the event reaches the Azure Event Hub

Here in Azure Stream Analytics this is represented as EventEnqueuedUtcTime

Azure Stream Analytics

Then the event reaches Azure Stream Analytics

In order for Azure Stream Analytics to understand the events that are coming in, it creates a watermark just to understand the ordering of events

Why does it need to do this?

Clock Skews

There is no guarantee that all clocks for all systems are sychronized

Network latency

Delays for the events to reach the intended destination

Azure SQL database

Azure Event Hub

Azure Stream Analytics

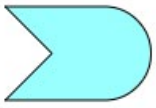
Hence you could have late arriving events or early arriving events

Sometimes events can be parsed, wherein they don't get generated that frequently

Or because there are too many partitions on the input side , and the data takes time to be spread across the partitions.

This is all things that Azure Stream Analytics needs to take care of to ensure that it gives the desired ordered state of results

Azure Stream Analytics - More on the time aspect



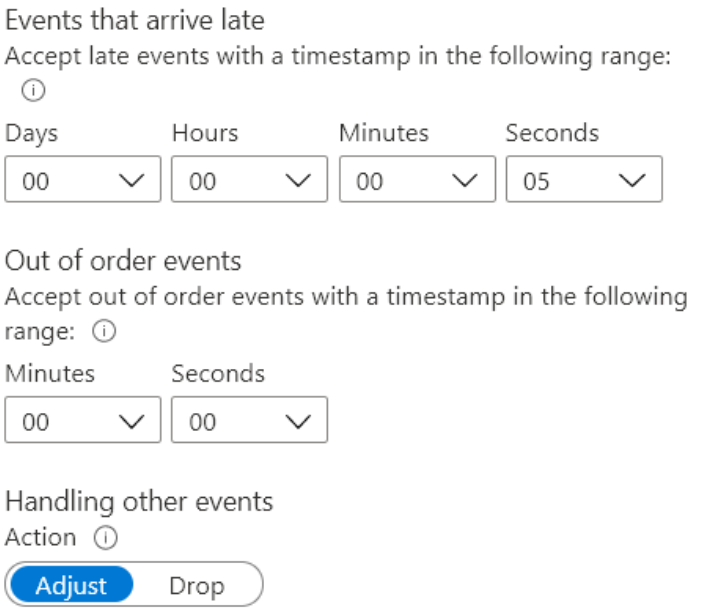
Azure Stream Analytics has received an event

10:00

The watermark is the largest event time - any out-of-order tolerance window size

If there is no incoming event, then the watermark is the current estimated arrival time minus the late arrival tolerance window

With the below settings you can decide either to adjust the time of the event or drop the events



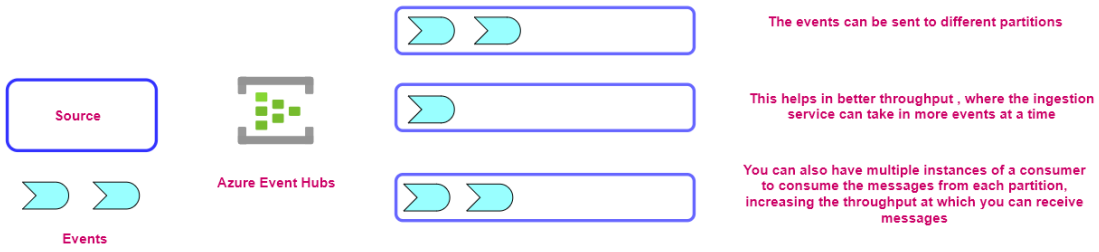
In case the events time is adjusted , it will adjust the System.Timestamp value and not the event time field.

Late Arrival policy - 15 seconds

Event No.	Event Time	Arrival Time	System.Timestamp	Explanation
1	00:10:00	00:10:40	00:10:25	Event arrived late and outside tolerance level. So event time gets adjusted to maximum late arrival tolerance.
2	00:10:30	00:10:41	00:10:30	Event arrived late but within tolerance level. So event time does not get adjusted.
3	00:10:42	00:10:42	00:10:42	Event arrived on time. No adjustment needed.

https://docs.microsoft.com/en-us/azure/stream-analytics/event-ordering

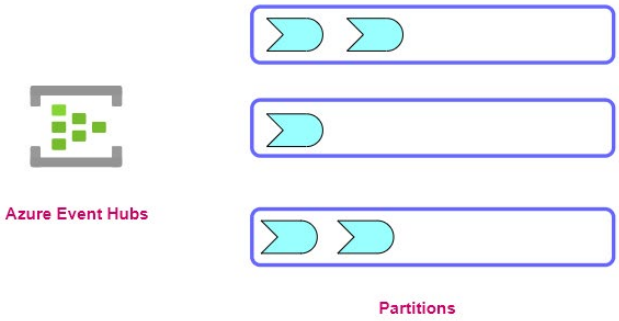
Azure Event Hubs and Stream Analytics – Partitions



When sending events from the source, you can decide which attribute of your data can be used as a partition key

Like a per-device or user identity attribute

Then the events can be split across the multiple partitions



The job can also work on the different partitions in parallel

This helps to increase the efficiency of processing

So Azure Stream Analytics can make use the partitions that are available in Azure Event Hubs

For Compatibility level 1.2

Azure Stream Analytics can already make use of partitions based on the partitions of the events in Azure Event Hub

For Compatibility level 1.0 or 1.1

Here you need to make explicit use of the partition key in the query

```
SELECT Id
FROM Input1 PARTITION BY PartitionId
```

Sometimes you might want to explicitly repartition your input if you don't have control over the partition key

You can also specify the number of output partitions especially if your destination is also Azure Event Hubs

In such a case always remember to ensure the number of input and output partitions are the same if you are using Azure Event Hubs for both the input and output

Calculating the maximum number of streaming units

Calculate the max streaming units for a job

All non-partitioned steps together can scale up to six streaming units (SUs) for a Stream Analytics job. In addition to this, you can add 6 SUs for each partition in a partitioned step. You can see some **examples** in the table below.

Query	Max SUs for the job
• The query contains one step. • The step is not partitioned.	6
• The input data stream is partitioned by 16. • The query contains one step. • The step is partitioned.	96 (6 * 16 partitions)
• The query contains two steps. • Neither of the steps is partitioned.	6
• The input data stream is partitioned by 3. • The query contains two steps. The input step is partitioned and the second step is not. • The **SELECT** statement reads from the partitioned	24 (18 for partitioned steps + 6 for non-partitioned steps

Azure Event Hubs - High Availability

High Availability for Azure Event Hubs

By default Azure Event Hubs can withstand if the underlying individual machines goes down

You can also enable Availability Zone support for Azure Event Hubs

But what happens if the entire region goes down

North Europe

West Europe

Create a pairing between both Event Hubs

If you are sending the events to an Azure storage account

And if you want high availability

Consider geo-replicated storage accounts