Contents

[DATA CLASSIFICATION 2](#_Toc126827740)

[DATA FORMAT 2](#_Toc126827741)

[TERMINOLOGIES 2](#_Toc126827742)

[DATA LAKE 2](#_Toc126827743)

[AZURE DATA LAKE STORAGE ACCOUNT 3](#_Toc126827744)

[CREATING DATA LAKE GEN-2 STORAGE ACCOUNT 3](#_Toc126827745)

[CREATING CONTAINIER AND UPLOAD BLOB OBJECTS 4](#_Toc126827746)

[CONNECTING STORAGE EXPLORER WITH DATA LAKE STORAGE ACCOUNT 5](#_Toc126827747)

[SQL 5](#_Toc126827748)

[AZURE SQL DATABASE 5](#_Toc126827749)

[ETL(EXTRACT-TRANSFORM-LOAD) 6](#_Toc126827750)

[AZURE DATA FACTORY 6](#_Toc126827751)

[WHAT PROBLEM DOES IT SOLVE? 6](#_Toc126827752)

[USAGE SCENARIOS 7](#_Toc126827753)

[COMPONENTS OF AZURE DATA FACTORY 9](#_Toc126827754)

[PIPELINE 9](#_Toc126827755)

[ACTIVITIES 9](#_Toc126827756)

[DATASET 10](#_Toc126827757)

[LINKED SERVICE 11](#_Toc126827758)

[MAPPING DATA FLOWS 12](#_Toc126827759)

[INTEGRATION RUNTIMES 12](#_Toc126827760)

[ADDITIONAL CONCEPTS 13](#_Toc126827761)

[FLOW PROCESS OF AZURE DATA FACTORY 13](#_Toc126827762)

[INPUT DATASET 14](#_Toc126827763)

[PIPELINE 14](#_Toc126827764)

[OUTPUT DATASET 14](#_Toc126827765)

[LINKED SERVICE 15](#_Toc126827766)

[GATEWAY 15](#_Toc126827767)

[CLOUD 15](#_Toc126827768)

# DATA CLASSIFICATION

|  |  |
| --- | --- |
| STRUCTURED DATA | * Tabular Data that is represented by rows and columns in a DB |
| SEMI-STRUCTURED DATA | * JSON * Key value data store- The values are based on the key. The key value data store   can be stored also in the form of tables, but over here, we don't have any sort of relationships between the tables as we have in a SQL based database system. |
| UNSTRUCTURED DATA | Data like having audio files, video files, images, etc. All of these are binary objects. |

## DATA FORMAT

|  |  |
| --- | --- |
| JSON | * JavaScript Object |
| AVRO | * This is row-based file format * Each record in the file contains a header that describes the structure of the data in the record * Stores data in compressed format |
| CSV | Comma Separated Values |
| PARQUET | Unlike row-based formats such as CSV or Avro, Apache Parquet is column-oriented – meaning the values of each table column are stored next to each other, rather than those of each record |

# TERMINOLOGIES

|  |  |
| --- | --- |
| DTU | * Database Transaction Unit. This unit is a combined measure of CPU, memory, IOPS, etc. |

# DATA LAKE

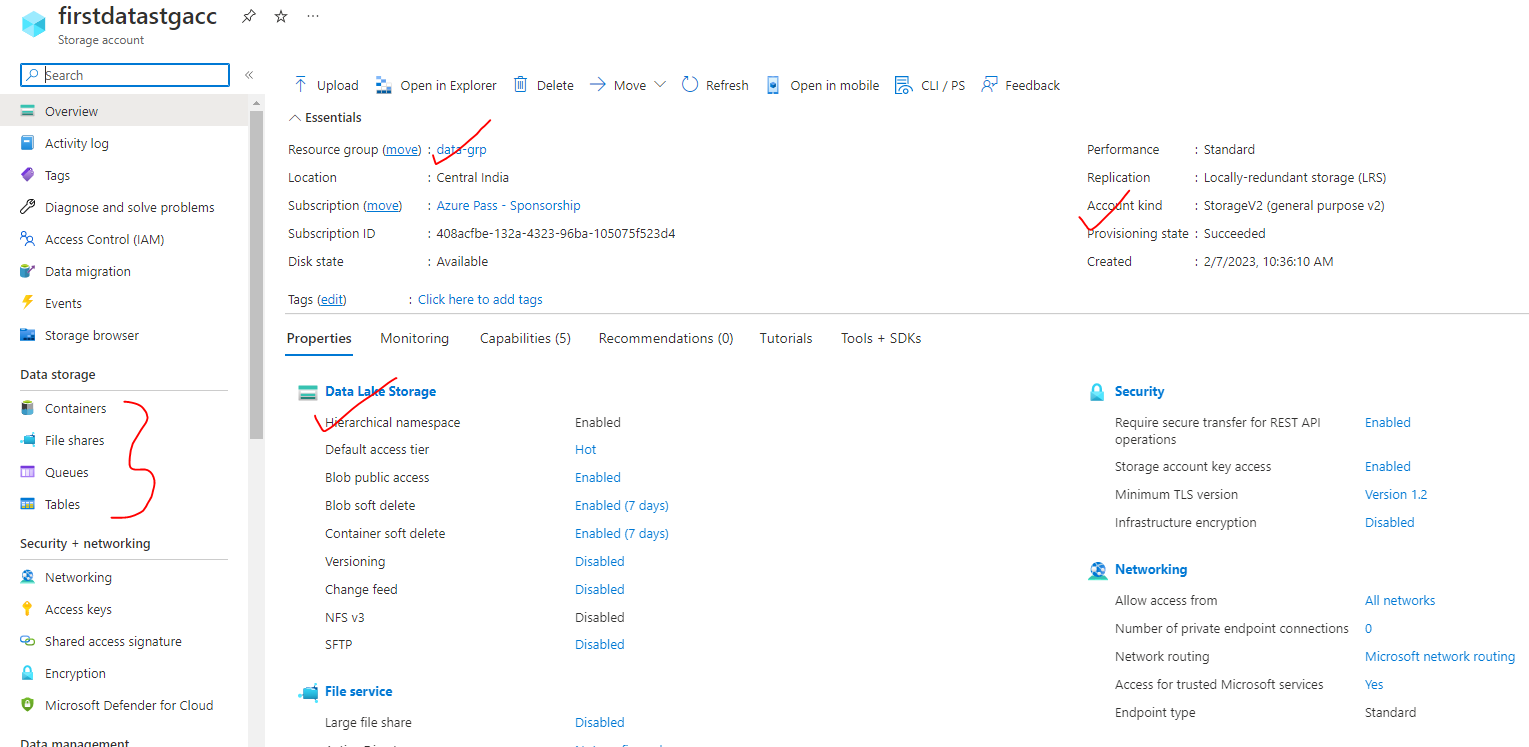
|  |  |
| --- | --- |
| **DATA LAKE IN AZURE**  Data lakes - Azure Architecture Center | Microsoft Learn | * A data lake is a system or [repository of data](https://en.wikipedia.org/wiki/Data_repository) stored in its natural/raw format(images, text, JSON format), usually object [blobs](https://en.wikipedia.org/wiki/Binary_large_object) or files. * Data lake is used when we are working with large data sets and data is arriving in large volumes and at fast rate. * The data lake has been provided as PAAS in Azure using Azure data lake storage account. |

# AZURE DATA LAKE STORAGE ACCOUNT

* Azure Data Lake Gen2 storage accounts is a service that provides the option of hosting a data lake on Azure.
* Data lake account can be used when we are working with large data sets and data is arriving in large volumes and at fast rate.
* ***Data Lake Storage account is built on top of Azure Blob Storage. Here we make use of features of a hierarchical name space on top of Azure block storage itself. This hierarchy helps to organize the objects and the files into a hierarchy of directories for efficient data access.***
* Data lake is used to store large amount of data in its native, raw format. It is optimized for storing terabytes and petabytes of data

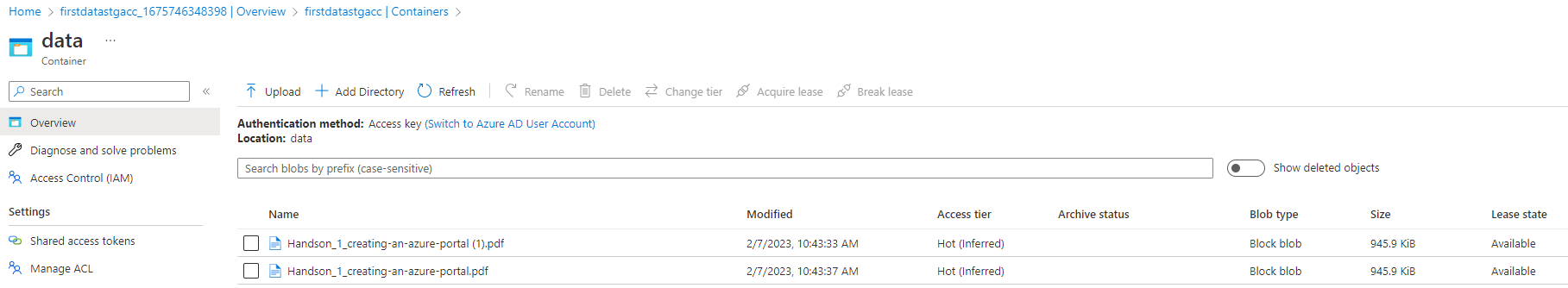
## CREATING DATA LAKE GEN-2 STORAGE ACCOUNT

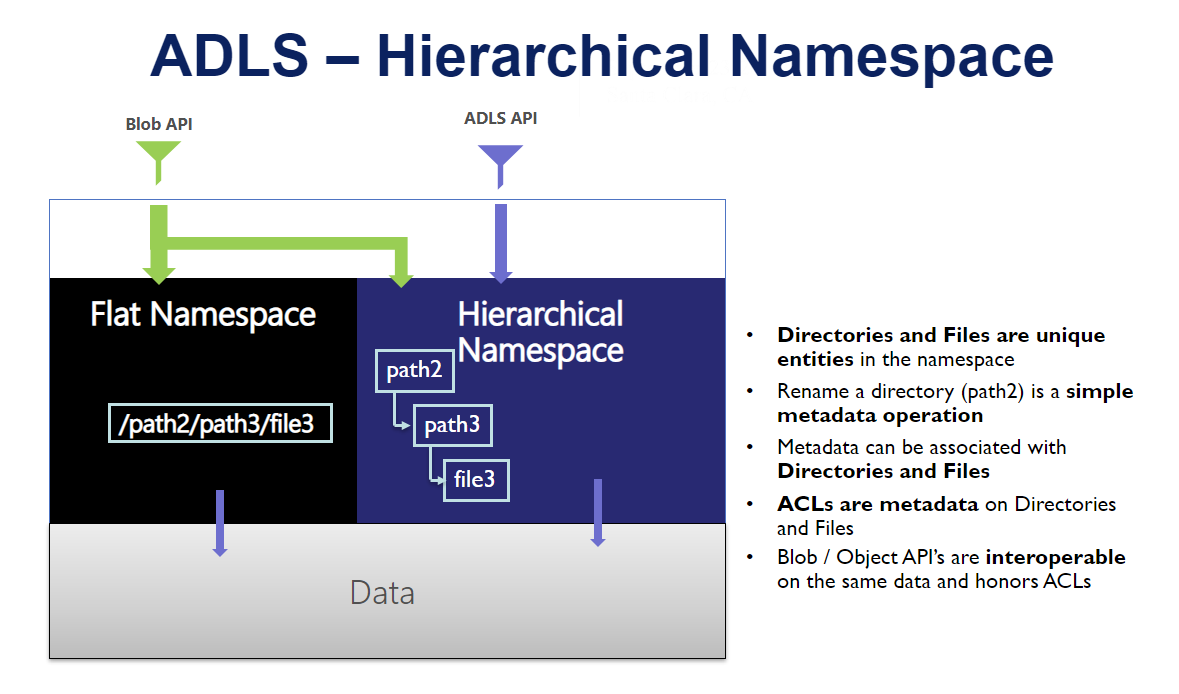
|  |  |
| --- | --- |
|  | * The data lake storage account is created same as storage account -with a difference |



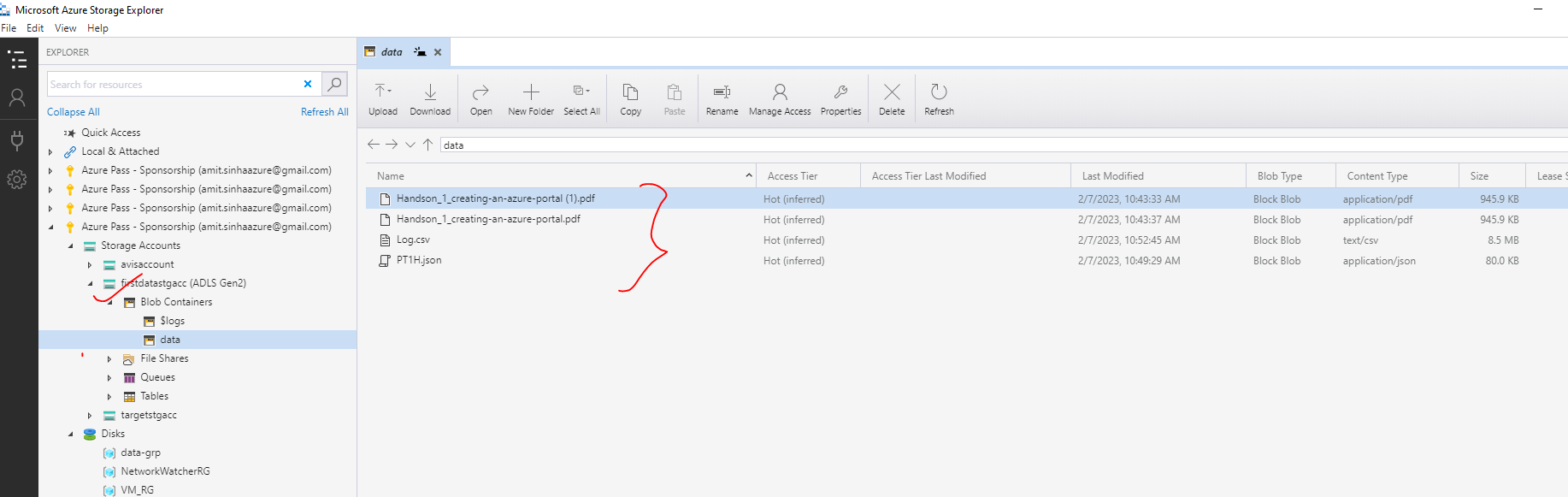
### CREATING CONTAINIER AND UPLOAD BLOB OBJECTS







## CONNECTING STORAGE EXPLORER WITH DATA LAKE STORAGE ACCOUNT



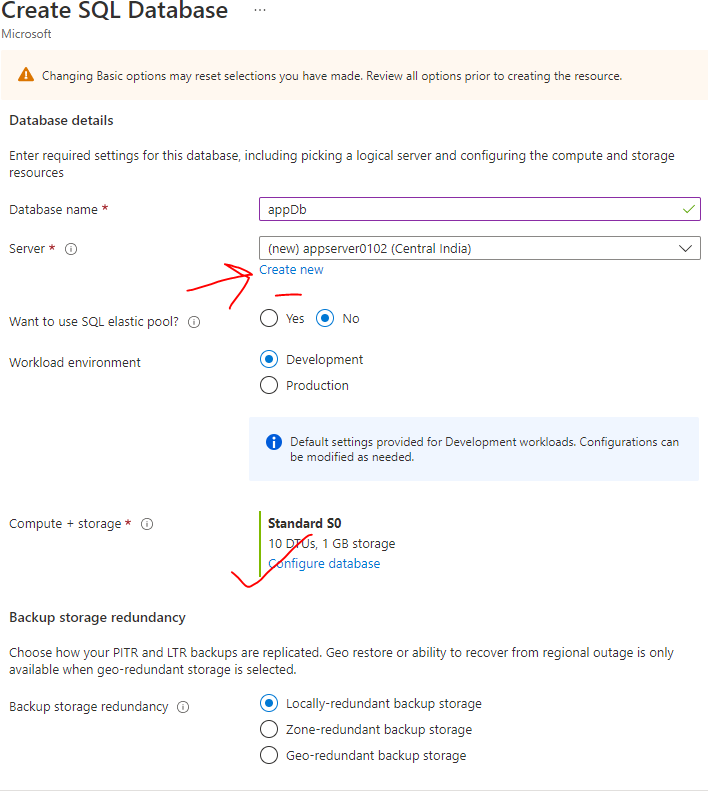
The data lake storage account can be connected to storage explorer- same as General Purpose Storage account. The different ways to connect are (refer the details in Azure Storage Account documentation)

1. **SUBSCRIPTION**
2. **ACCESS KEYS**
3. **SHARED ACCESS SIGNATURE**

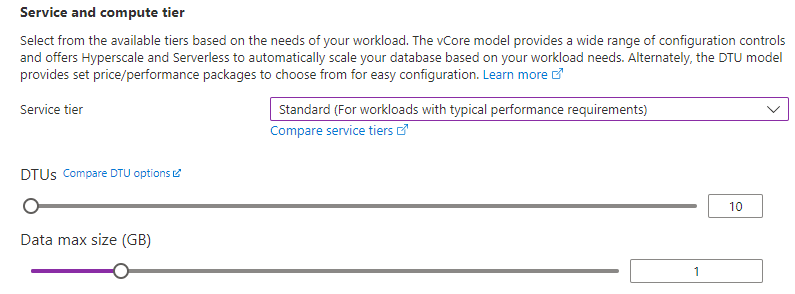
# SQL

## AZURE SQL DATABASE

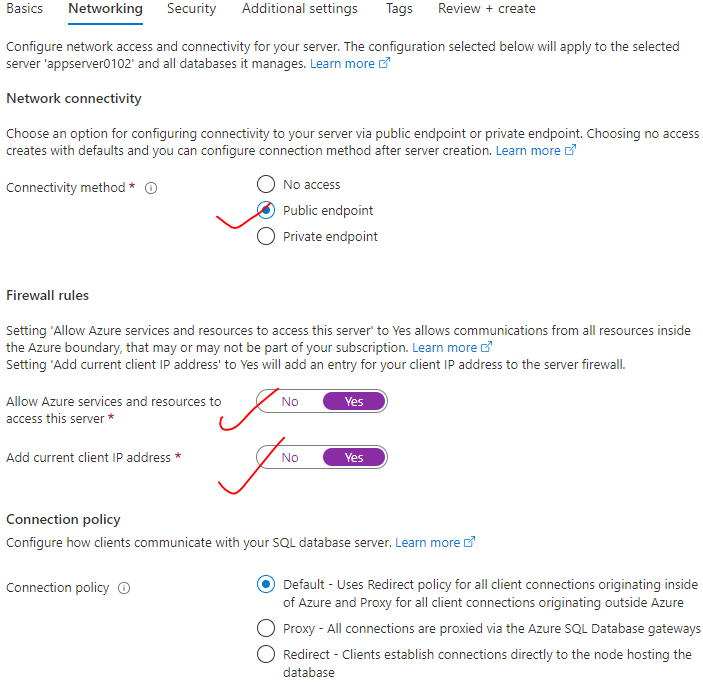
* It’s a PAAS in Azure. To create SQL database server - Select SQL database service 🡪 Create SQL Database.



### CONFIGURING THE DATABASE



### NETWORKING



### ADDITIONAL SETTINGS

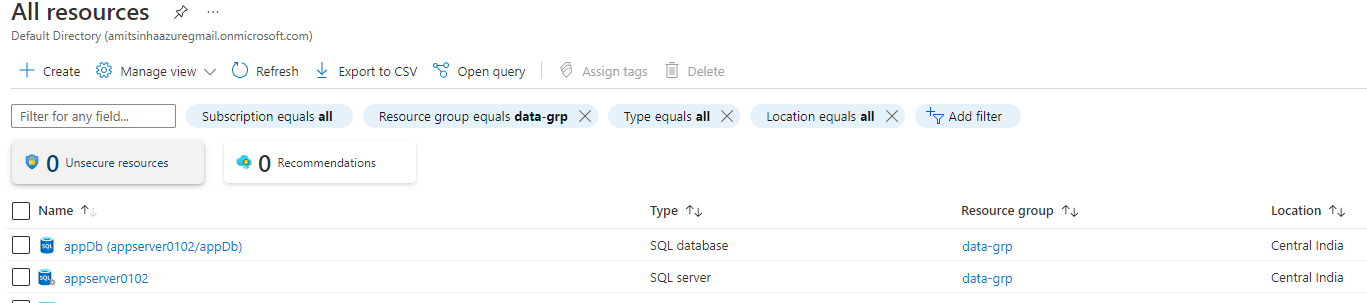
* We can make use of sample data as well. This will pre-populate the DB with tables with sample data

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

* Note - This implementation creates two resources.

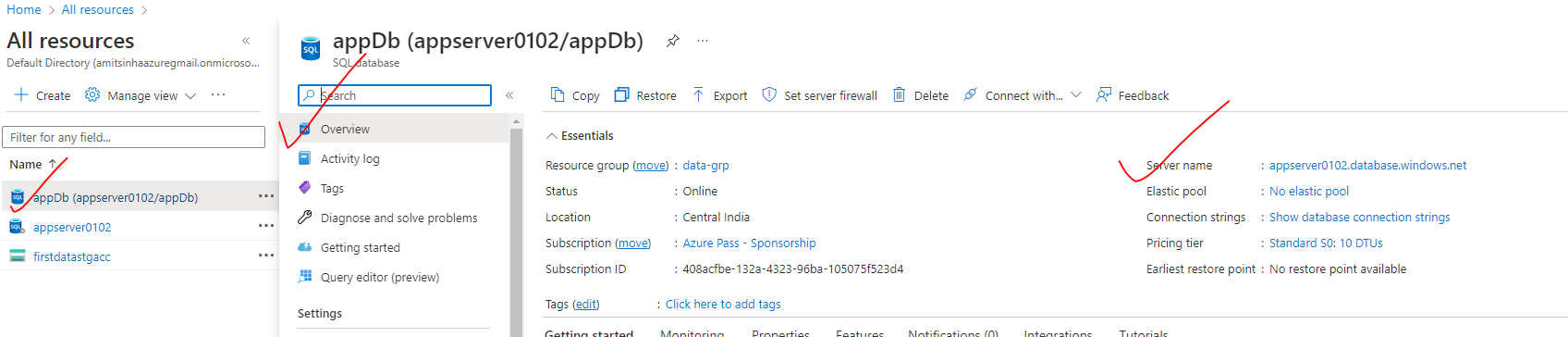
1. **AZURE SQL DATABASE SERVER**
2. **AZURE SQL DATABASE.**

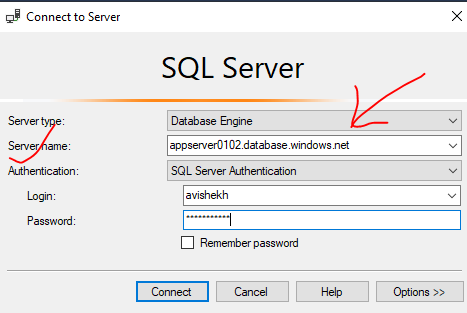
* **We can have one DB server with multiple SQL data base.**



### CONFIGURING SQL SERVER STUDIO

* To connect the SQL server studio with SQL Database server and hence access the SQL DB



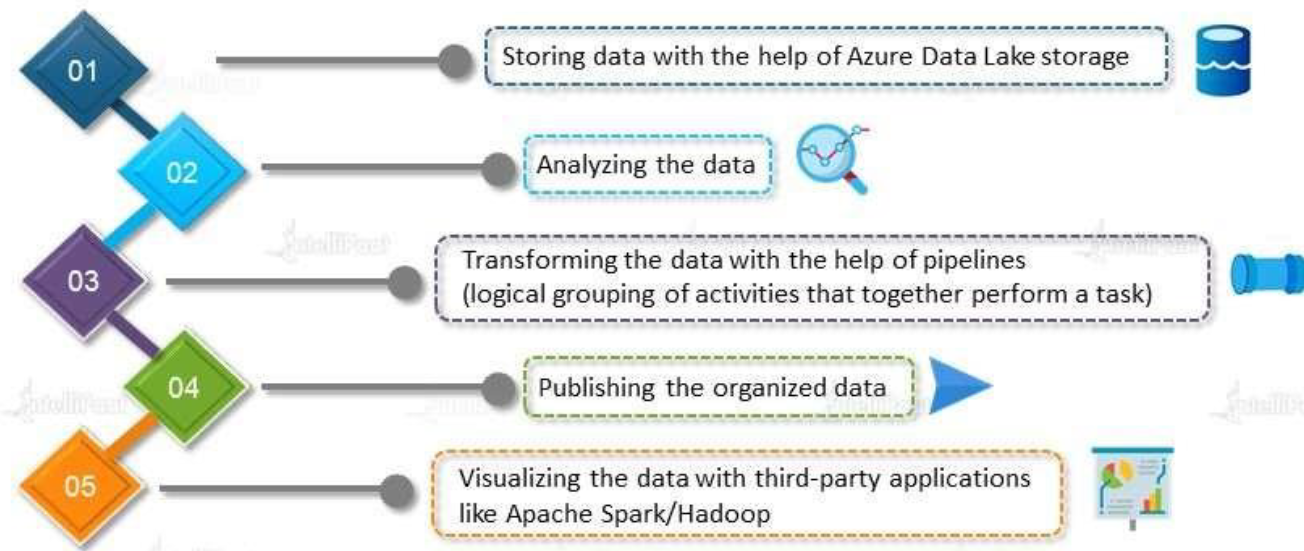


# ETL(EXTRACT-TRANSFORM-LOAD)

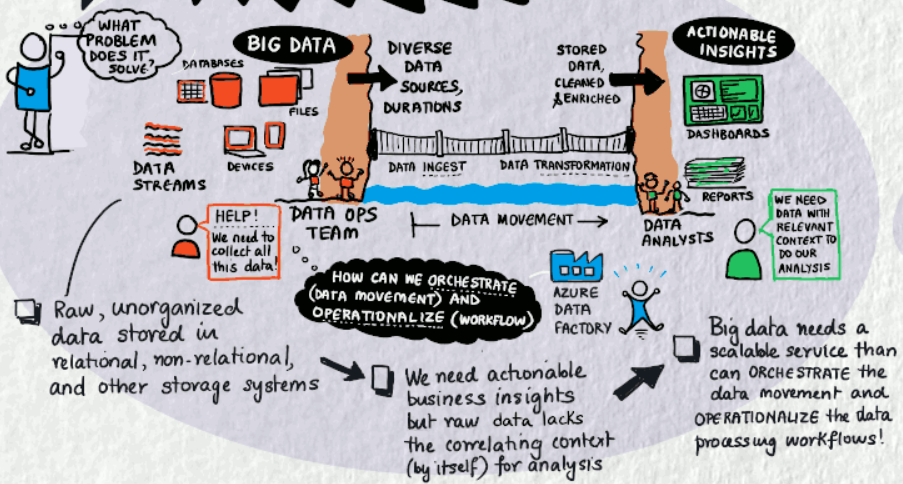
|  |  |
| --- | --- |
| What is ETL? (Extract Transform Load) | Informatica | ETL is a process of extracting data from a variety of data sources and load the data into a target area. Example -extracting data from an Azure SQL Database or SQL server instance hosted in an on-premise environment and load it on to an Azure SQL Data Warehouse after perform some sort of transformation of the data |

# AZURE DATA FACTORY

|  |  |
| --- | --- |
|  | * In the world of big data, raw, unorganized data is often stored in relational, non-relational, and other storage systems. However, on its own, raw data doesn't have the proper context or meaning to provide meaningful insights to analysts, data scientists, or business decision makers. * Big data requires a service that can orchestrate and operationalize processes to refine these enormous stores of raw data into actionable business insights. Azure Data Factory is a managed cloud service that's built for these complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects. |



## WHAT PROBLEM DOES IT SOLVE?



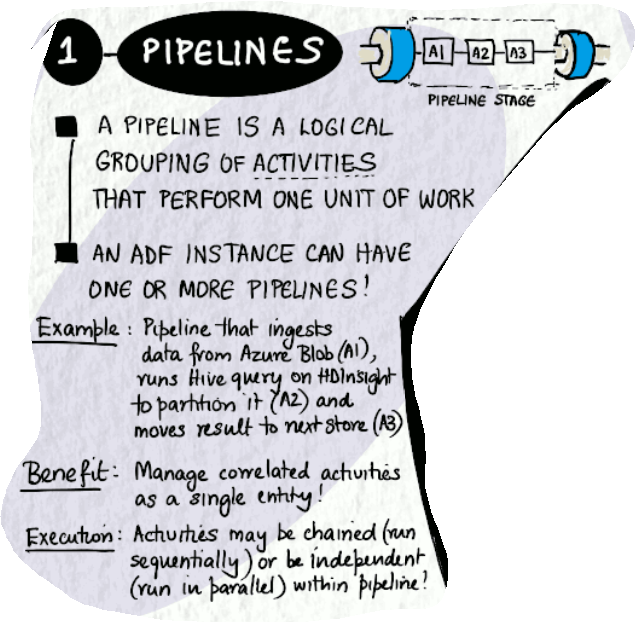
## USAGE SCENARIOS



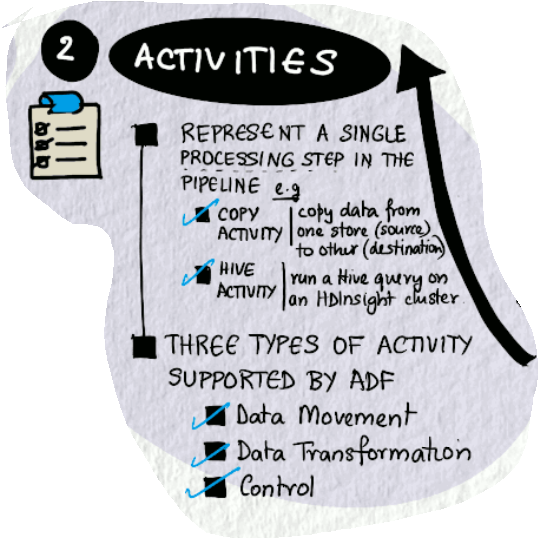
* For example, imagine a gaming company that collects petabytes of game logs that are produced by games in the cloud. The company wants to analyze these logs to gain insights into customer preferences, demographics, and usage behavior. It also wants to identify up-sell and cross-sell opportunities, develop compelling new features, drive business growth, and provide a better experience to its customers.
* To analyze these logs, the company needs to use reference data such as customer information, game information, and marketing campaign information that is in an on-premises data store. The company wants to utilize this data from the on-premises data store, combining it with additional log data that it has in a cloud data store.
* To extract insights, it hopes to process the joined data by using a Spark cluster in the cloud (Azure HDInsight), and publish the transformed data into a cloud data warehouse such as Azure Synapse Analytics to easily build a report on top of it. They want to automate this workflow and monitor and manage it on a daily schedule. They also want to execute it when files land in a blob store container.
* Azure Data Factory is the platform that solves such data scenarios. **It is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale**.
* Using Azure Data Factory, we can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores. We can build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database.
* Additionally, we can publish your transformed data to data stores such as Azure Synapse Analytics for business intelligence (BI) applications to consume. **Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores and data lakes for better business decisions**.

## COMPONENTS OF AZURE DATA FACTORY

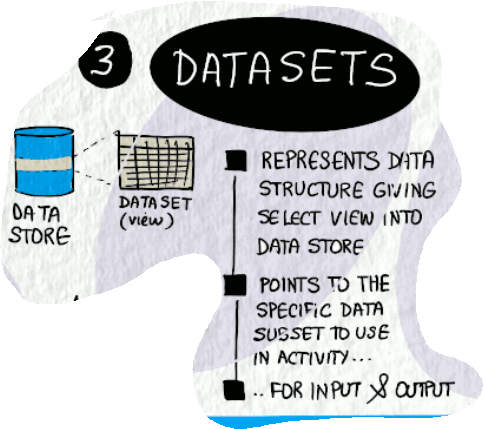
### PIPELINE



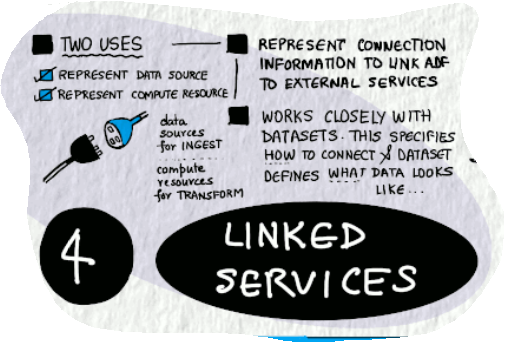
### ACTIVITIES



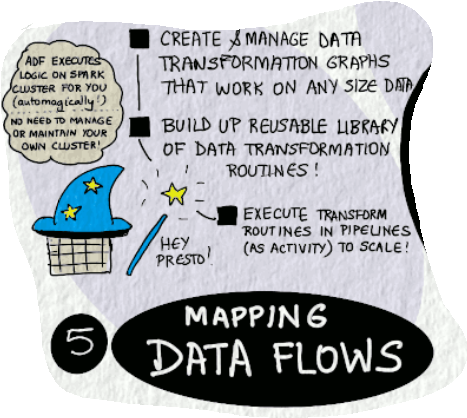
### DATASET



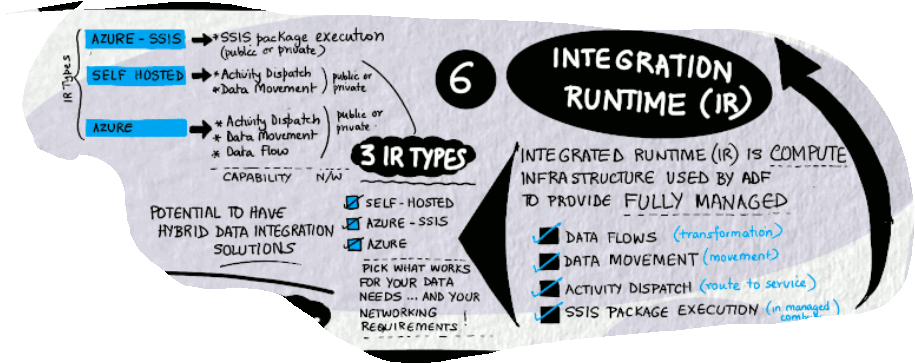
### LINKED SERVICE



### MAPPING DATA FLOWS

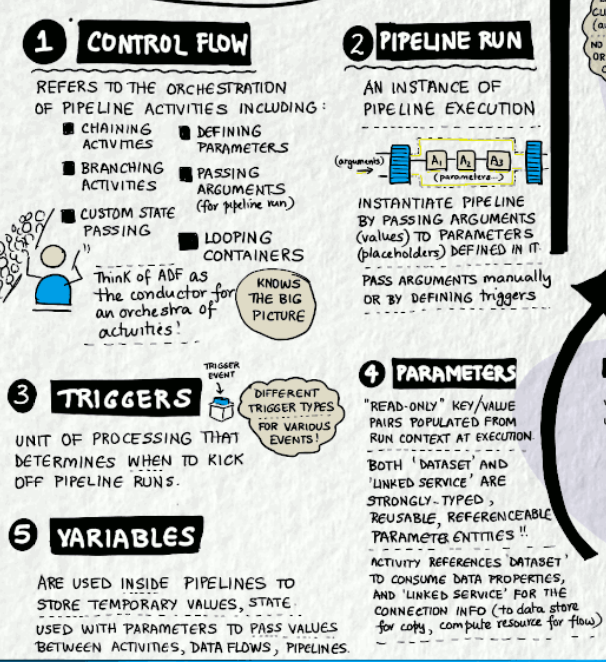


### INTEGRATION RUNTIMES



These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.

### ADDITIONAL CONCEPTS

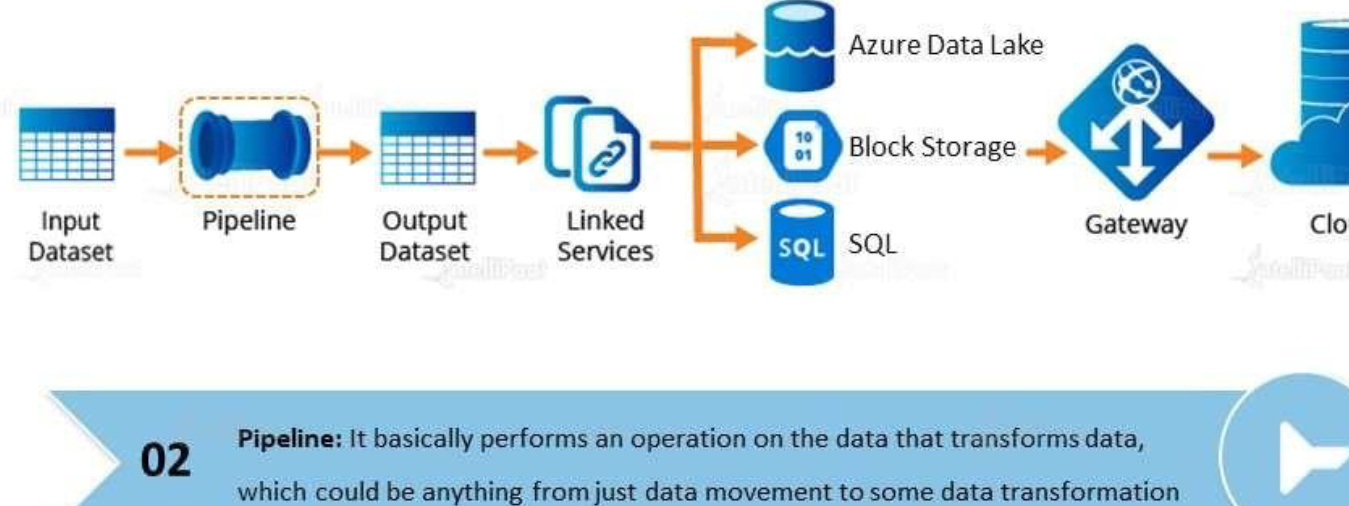


## FLOW PROCESS OF AZURE DATA FACTORY

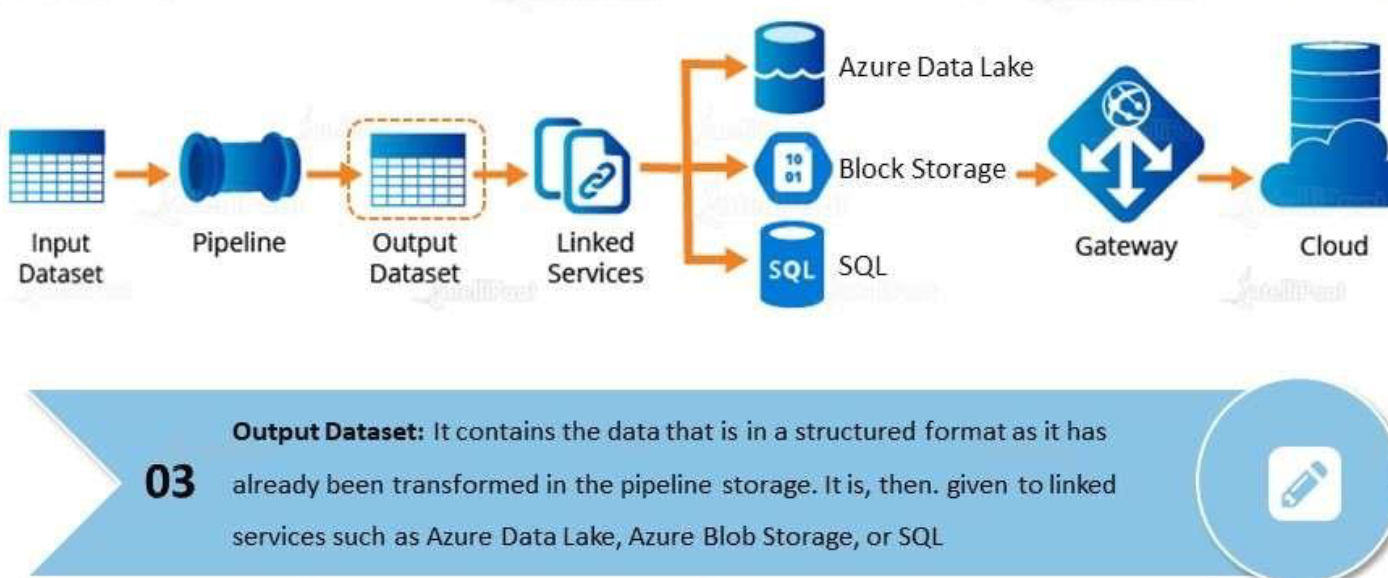
### INPUT DATASET



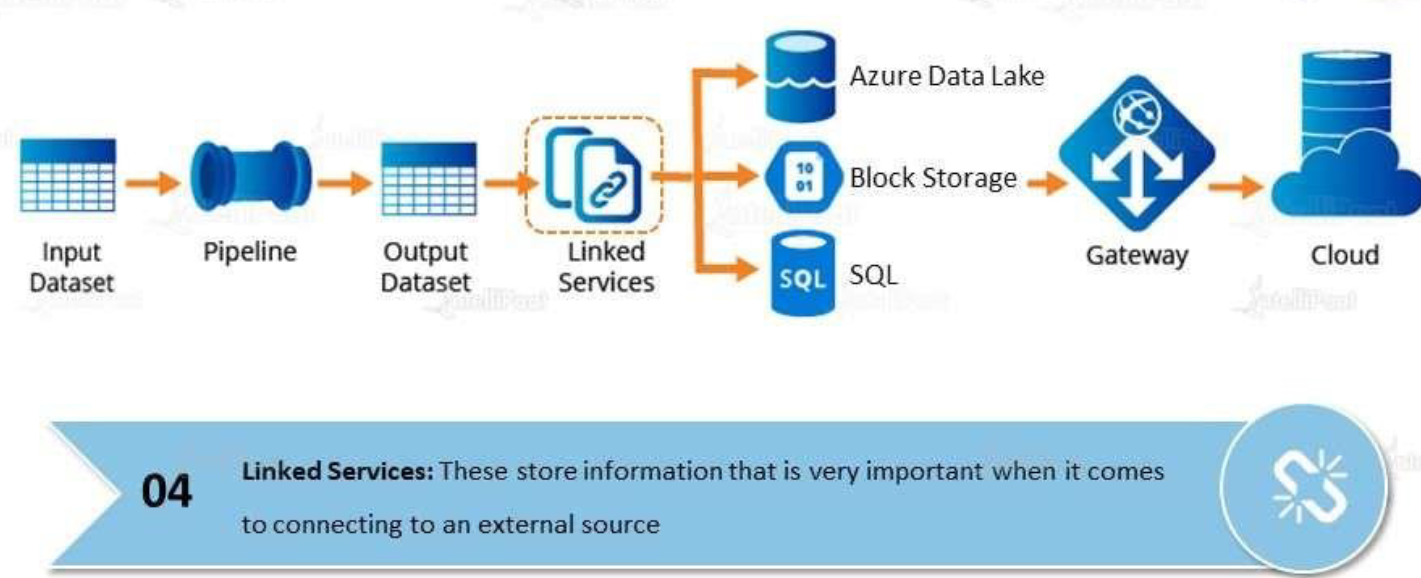
### PIPELINE



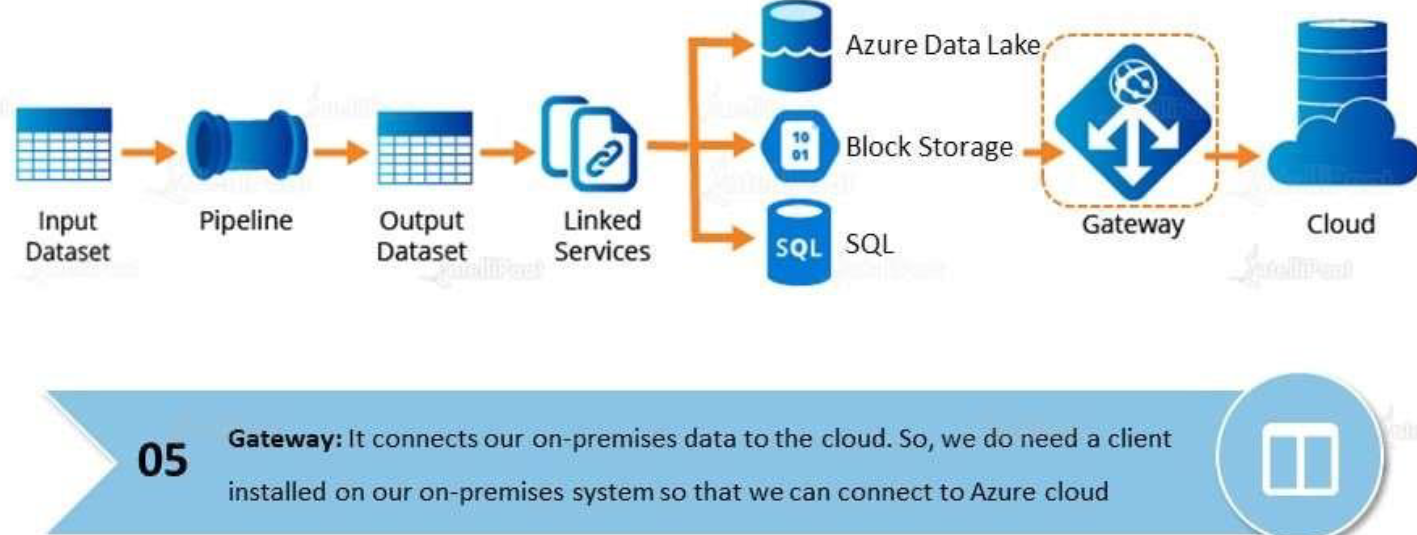
### OUTPUT DATASET



### LINKED SERVICE



### GATEWAY



### CLOUD

