Azure data factory is used for data Extraction or Loading as Transformation with ADF is costly, for transformation we use Databricks / Py Spark

Medallion architecture

* Bronze layer or Raw layer: This contains the exact copy of data from the sources, even incase of a failure or the sources getting corrupted we still have the data in the raw layer as backup
* Silver or curated layer:
  + In this layer we do some transformations like renaming the columns based on the naming conventions and then data type mapping, clearing null values
* Gold Layer:
  + Here we will perform aggregate functions, window functions, joins etc..
  + This is used to prepare data for our end users, like data scientists, data analysts and power bi analysts.

Data sources:

* SQL SERVER, REST APIs, SFTP

Target:

* ADLS, azure data lakes

**3 types of data**:  
1.Strucrured data:data which is in the form of rows and columns (sql tables,oracle tables,csv files)  
2.semi-structured :Data which is in the form of key value pairs/tags (json,xml,)  
3.unstructured Data: Data wont be having any form (pdf files,mp3,images)

•**Definition: A design pattern for data architecture that organizes data into a series of layers or "medallions," typically bronze, silver, and gold.**

•

•**Bronze Layer:**

**Rawdata ingestion.**

**Stores unprocessed data from various sources.**

**Often includes data in its original format (e.g., logs, files).**

•**Silver Layer:**

•**Processed and cleaned data.**

•**Data is transformed and enriched for analysis.**

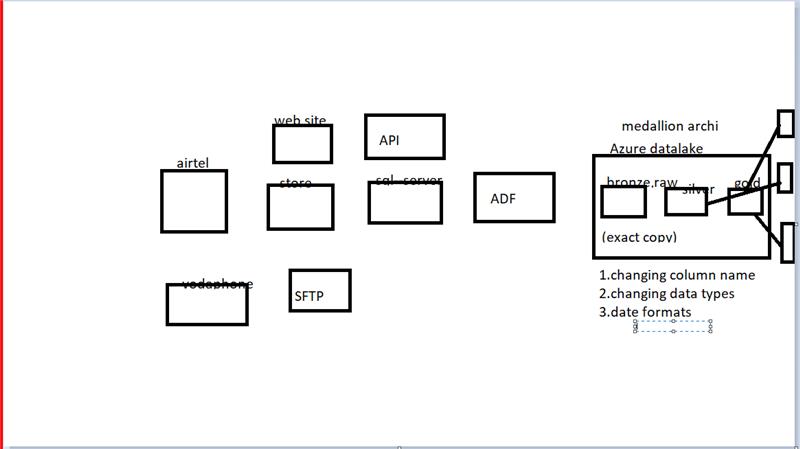
•**Includes structured and semi-structured data.**

•**Gold Layer:**

•**Curated, high-quality datasets.**

•**Optimized for analytics and reporting.**

•**Contains aggregated data ready for business intelligence (BI) tools.**



**difference Data Warehouse and Data Lake:**

Datawarehouse:  
1.it works on ETL flow(Extract transform and load)  
2.It is used to store for only Structured Data ( minimal semi-structured data)  
3.Time travel or version is difficult in DW

Data Lake:  
1. It works on ELT flow (extract load and transform)  
2. It can store all types of data (structured, semi-structured and unstructured data)  
3. Time travel or version is easily possible in DL(Delta-tables)

* We perform data preprocessing on the data present in raw/bronze layer and store this preprocessed data in silver layer, whereas the data in gold layer is stored after the silver layer data is aggregated or advanced data processing is applied.

Azure is a cloud provider and inside this we have a service called ADF (Azure Data Factory) which is a ETL tool.

And Azure data lake tool is used for storage.

Azure Key Vaults is used for security.

Azure Databricks for data processing.

Languages: python with spark, SQL, scala

1.Azure data Factory: extraction(ETL tool)  
key Componets :  
1.DataSet:It is a pointer /refernce which will be pointing source or sink(target)  
2.linked Service:it connection string which used to connect source or sink  
Note :Connection String :it holds the information like username, password, api links which help us to connect to ADF depending on source /sink  
3.Activties:its a logical entity which helps us to perform some tasks(copy activity each, wait)  
4.pipline:set activities will be kept in one pipeline and whenever we runs this pipeline   
5.intergration run time(IR):It provides the computation infrastructure to ADF from data movement / data migration  
There 3 types of IR  
1.Azure IR/Auto resolved integration runtime: it used to connect public to public networks(example data transfer between adls to adls)  
2.Self-hosted IR: it used to connect private to public network(example data transfer between sql server to ADLS)  
3.SSIS IR: This IR was built mainly for SSIS packages

Note1: Default IR is Azure IR/Auto resolved integration runtime

Note2: Self-hosted IR needs to install in our systems

6.Triggers:it used to schedule the piplines in productions

3 types  
1.Scheduled triggers  
2.Event based triggers  
3.tublimng - window triggers

▣Step 1: Navigate to the Azure Data Factory portal, and then select the "Manage" option located as the fourth option on the left side.

▣Step 2: Within the portal, locate and click on the "Integration Runtime" option available on the left.

▣Step 3: Click on the "New" option, choose "Self-hosted IR," and proceed by clicking on the "Continue" option.

▣Step 4: Provide a name for the Integration Runtime (IR) and, under Option 2 (Manual setup), select "Download and install integration runtime." Install the runtime on your local machine, choosing the latest version.

▣Step 5: Click on "Create," and make sure to copy the keys generated during this process and keep them accessible.

▣Step 6: Following the installation, input the keys copied in Step 4 and proceed by clicking on "Next."

▣Step 7: Upon successful completion, you will observe the Integration Runtime status as "Running" in Azure Data Factory (ADF).