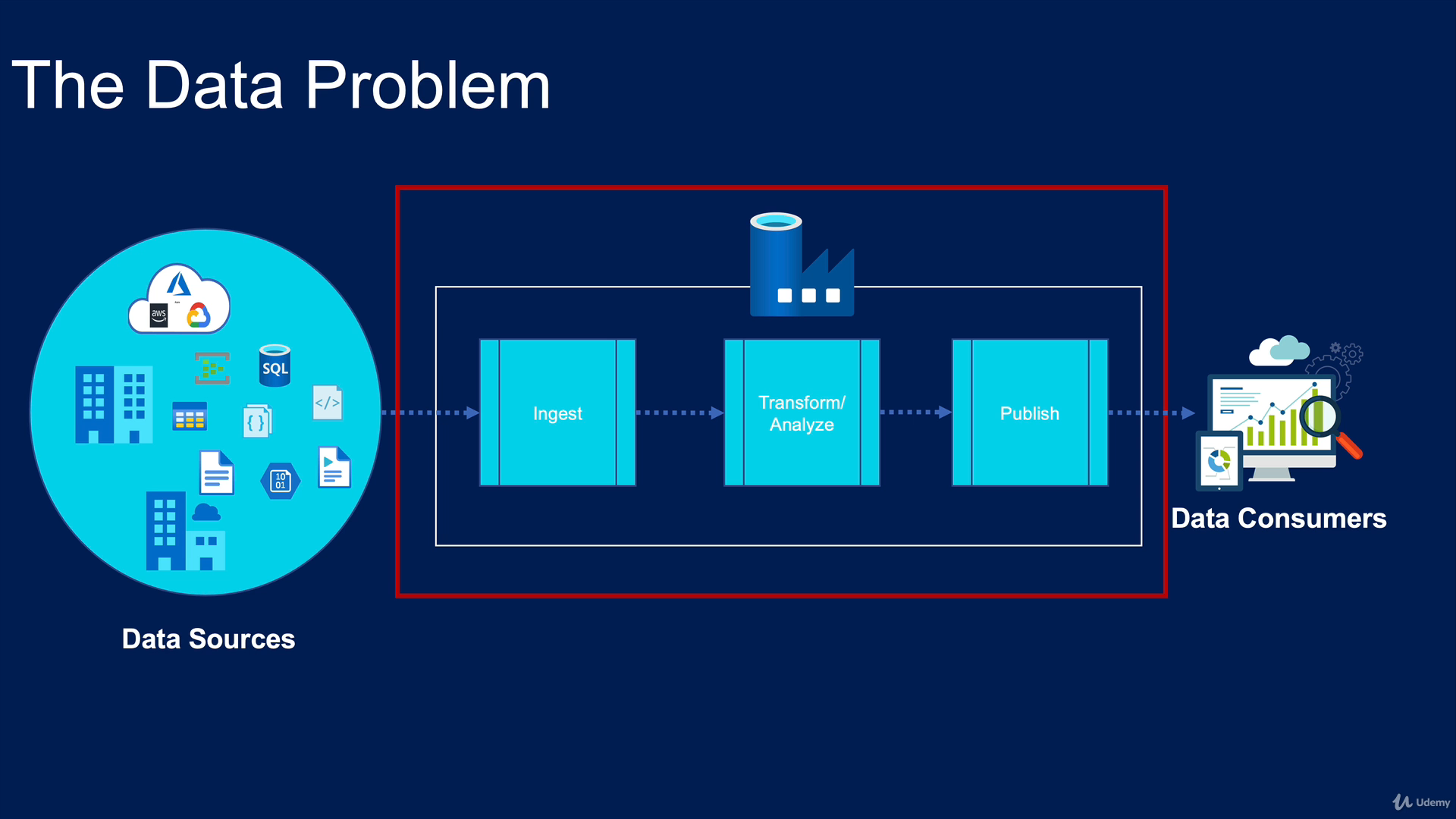
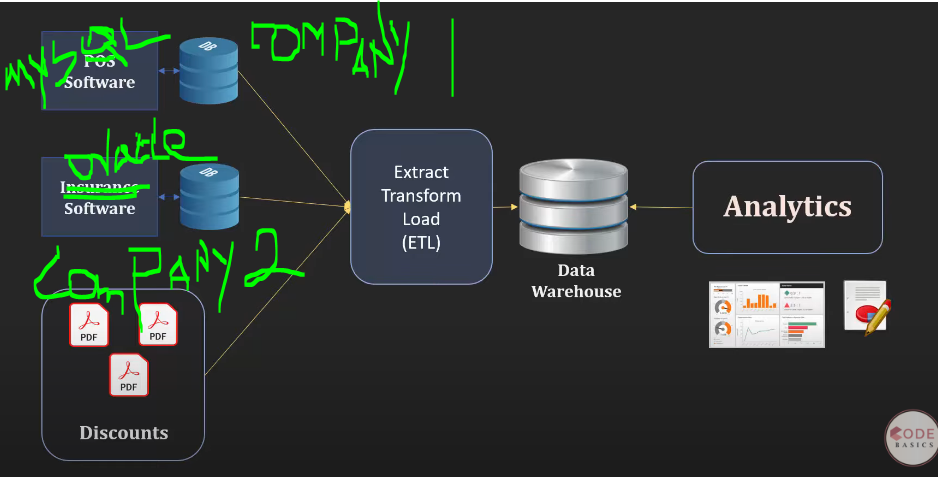
Azure Data Factory

Definition: A fully managed ,serverless data Integration solution for ingestion, preparing and transforming all your data at large scale.

Data Factory help to solve data problem by Ingest, Analyze, Publish to consumer through PowerBi.

ETL : Extract Data , Transform Data, Load Data in Data WareHouse, then from data warehouse Datascientist can use Data to Get Insights And Publish Reports Or Display in Dashboards.



Structured Data: Data present in Database Like Sql, Oracle, db2 in tables format.

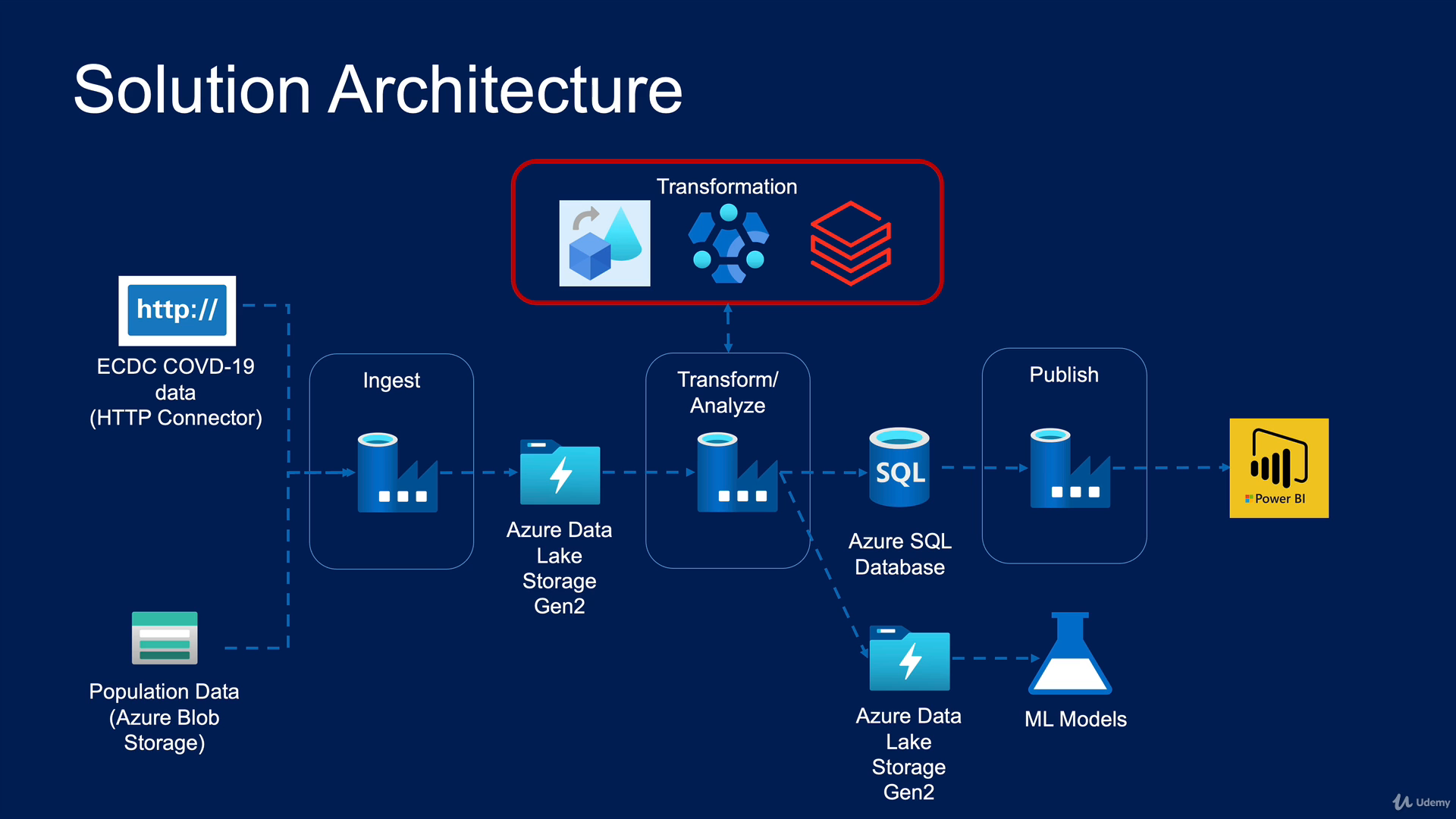
Untructured Data: Data stored in files/Forms/pdfs/Images etc.

Database: Database is where we stored data like SQl,Oracle etc. This Is majorly deigned for transactional and not to perform data Analytics.

DatawareHouse: DatawareHouse is also a database but it consumes the data from different databases and Optimized data to performing data analytics.

DataLake: here we tore both structured and unstructured Data, here we store all type of data, which we are not yet planned to perform DataAnalysis.

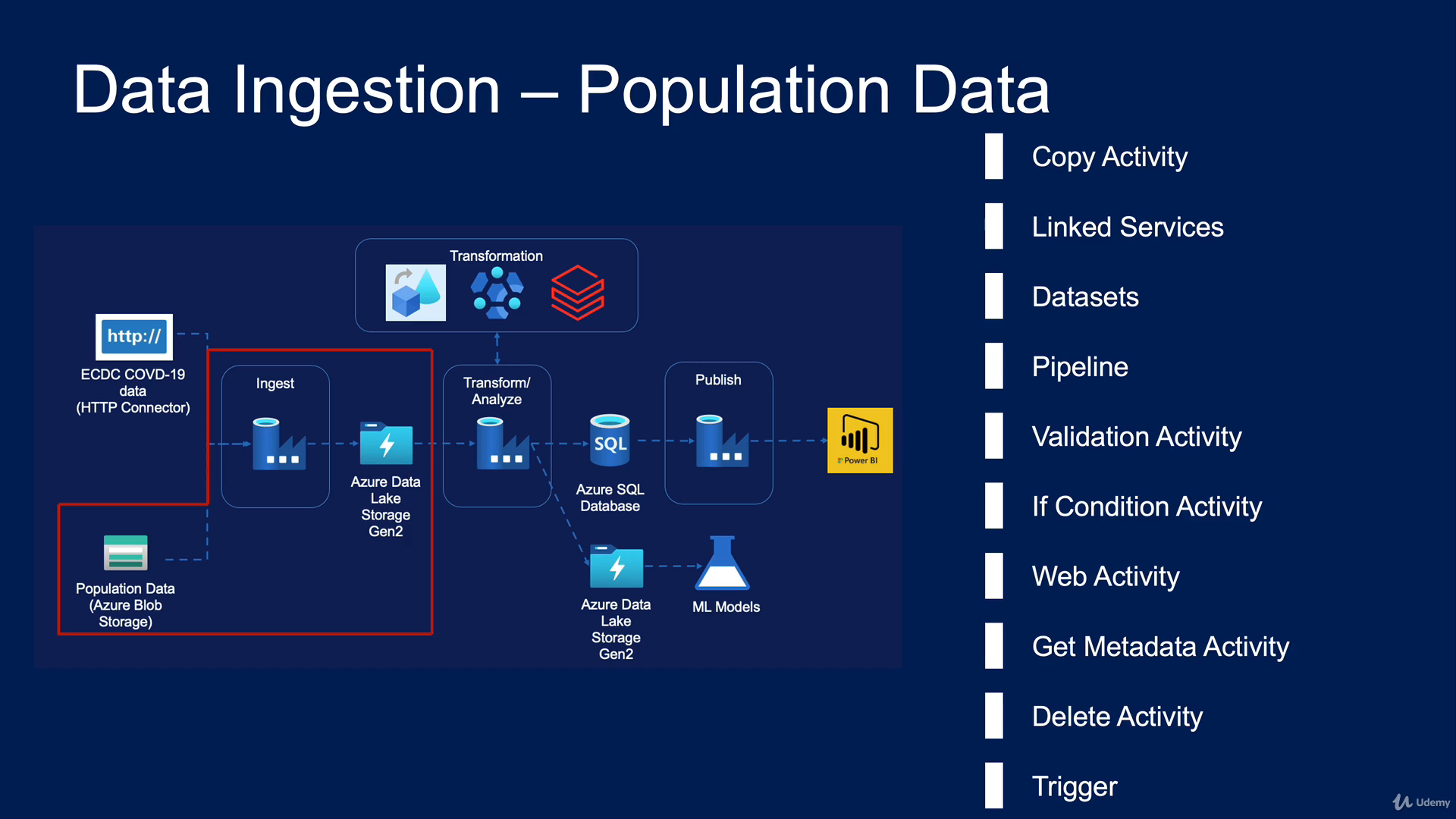
Creating COVID 19 Project

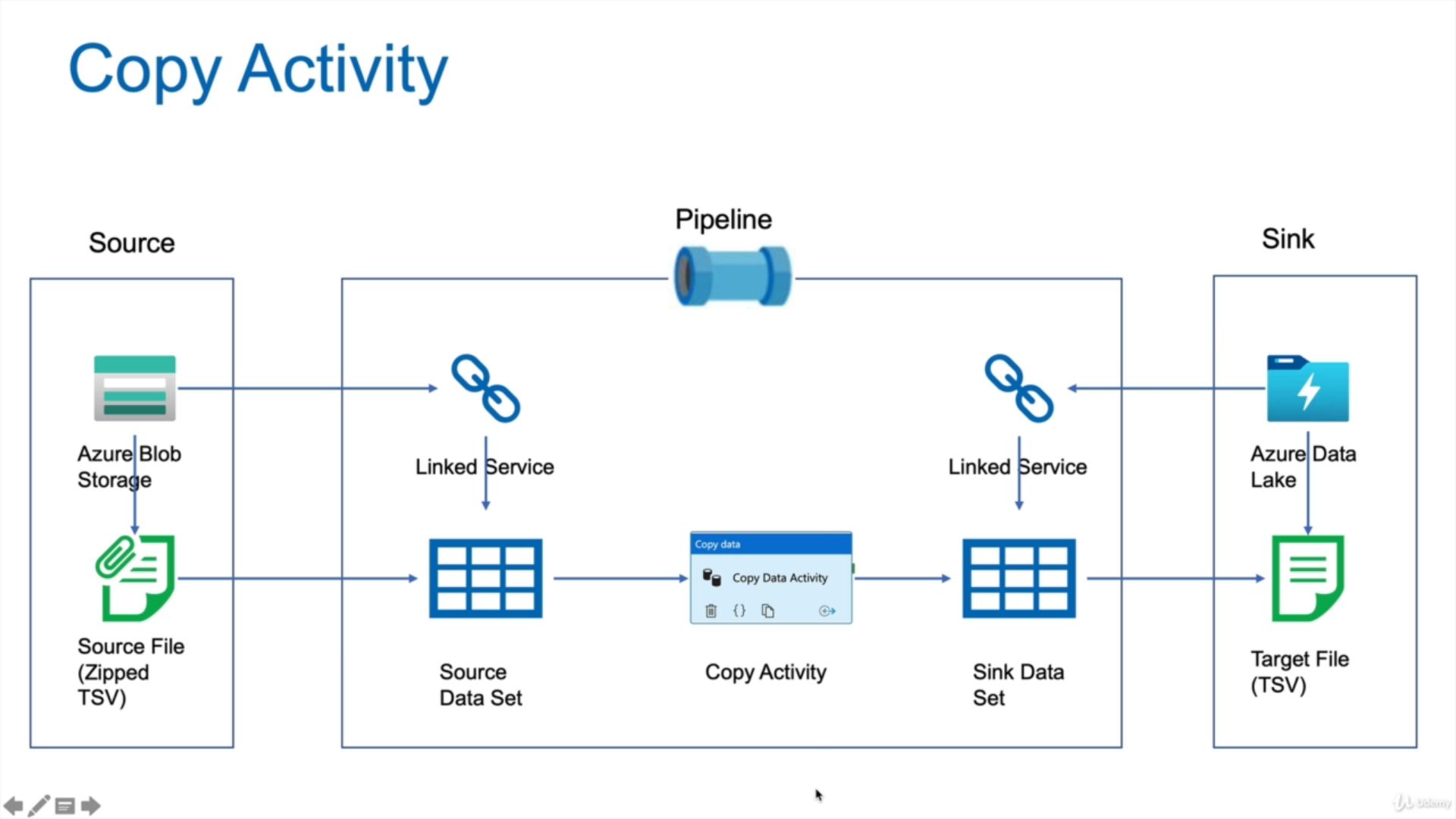


Steps:

1. Fetching Data from http Connector Or Also you can Fetch data from Azure Blob storage.
2. Ingest data and store into Azure Data Lake Storage.
3. Then Transform/ Analyze the data, transformation technique are Dataflow, HD Insights, Databrick. here you can use any 1 technique to transform the data from above transformation technique.
4. Then you pass Transform data in AzureData lake Storage for ML Models.
5. You can also store subset of data for power BI report generation in sql database.

Create different resources which needed in project in Azure Portal.





To copy from Blob storge we need Source Zip property and property of target file e.g .tsv.gz -> .tsv file so we do this by 1st converting zip file to source Dataset and store converted file to SinkDataset.

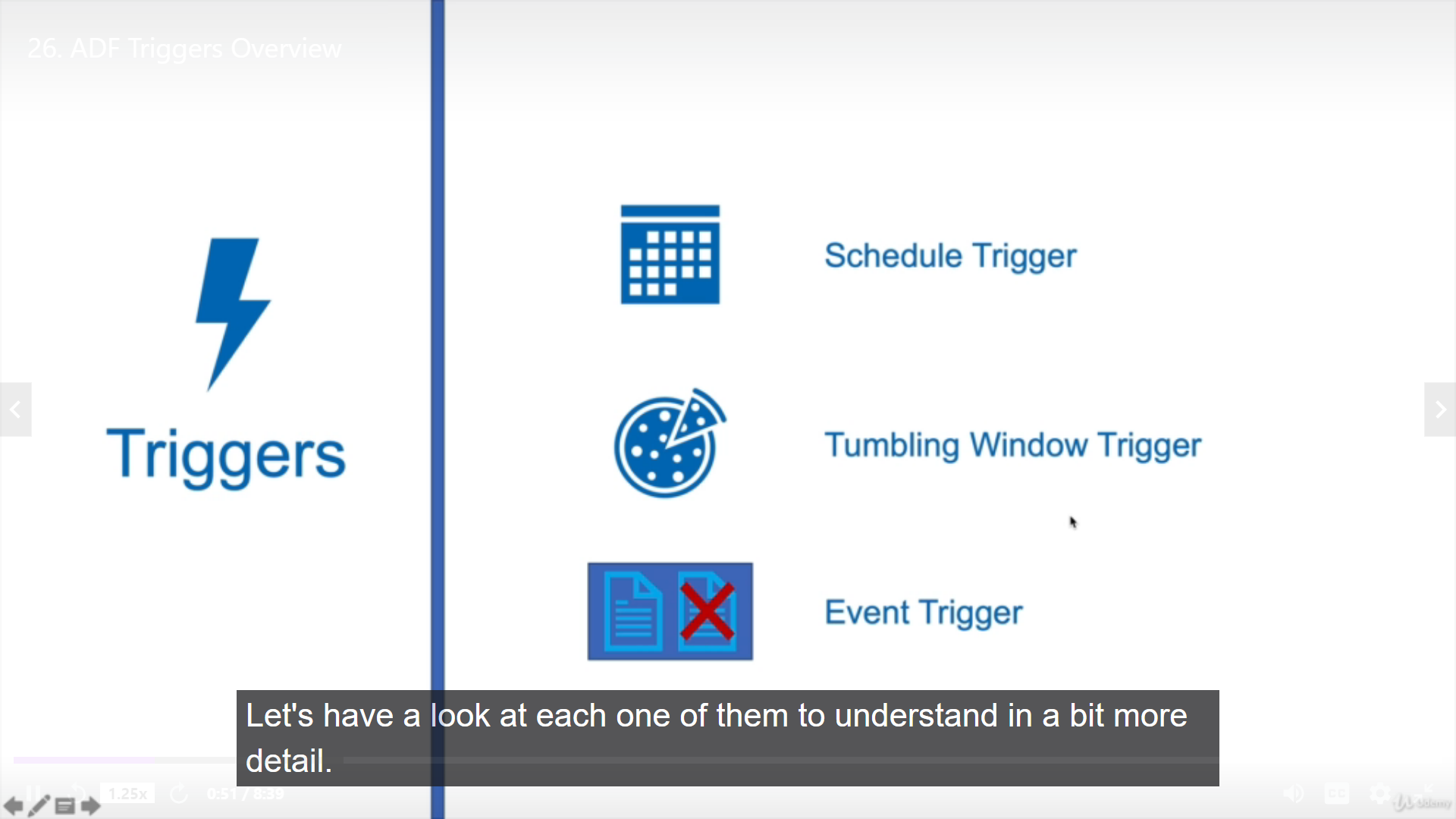
Source Dataset : as we need to transform the data rom csv to sequel table format, so this source datset will provide the meachanish and help to understand the data as per data required in output perspective.

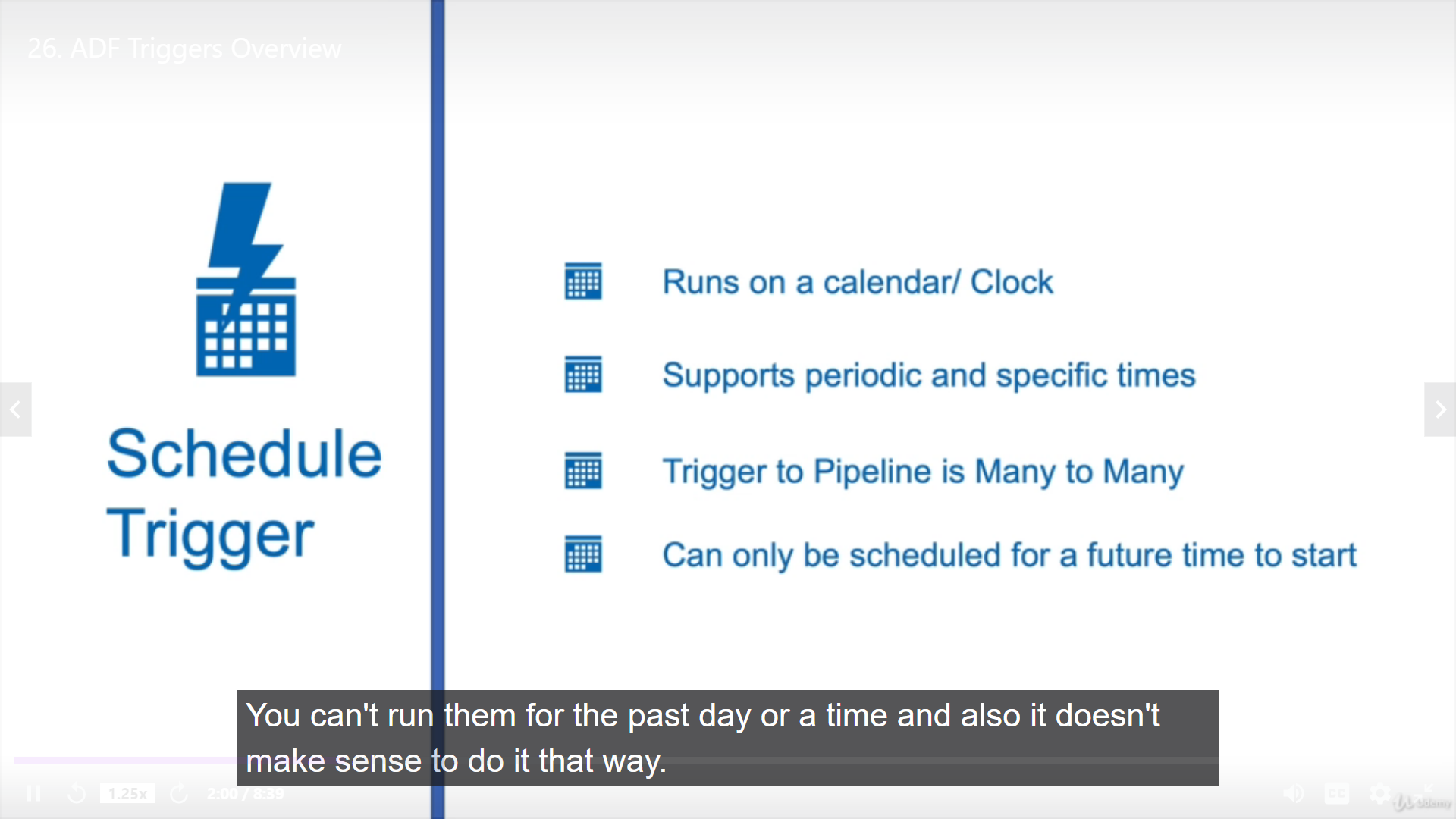
Blob Storage and Azure Datalake can connect through Link Service only.

And For Copying Activity does not done directly we need Pipeline for that.

After Completing Pipeline we can Schedule event like reading file after every 24 hour. This can help by Triggers.

There are 3 type of triggers

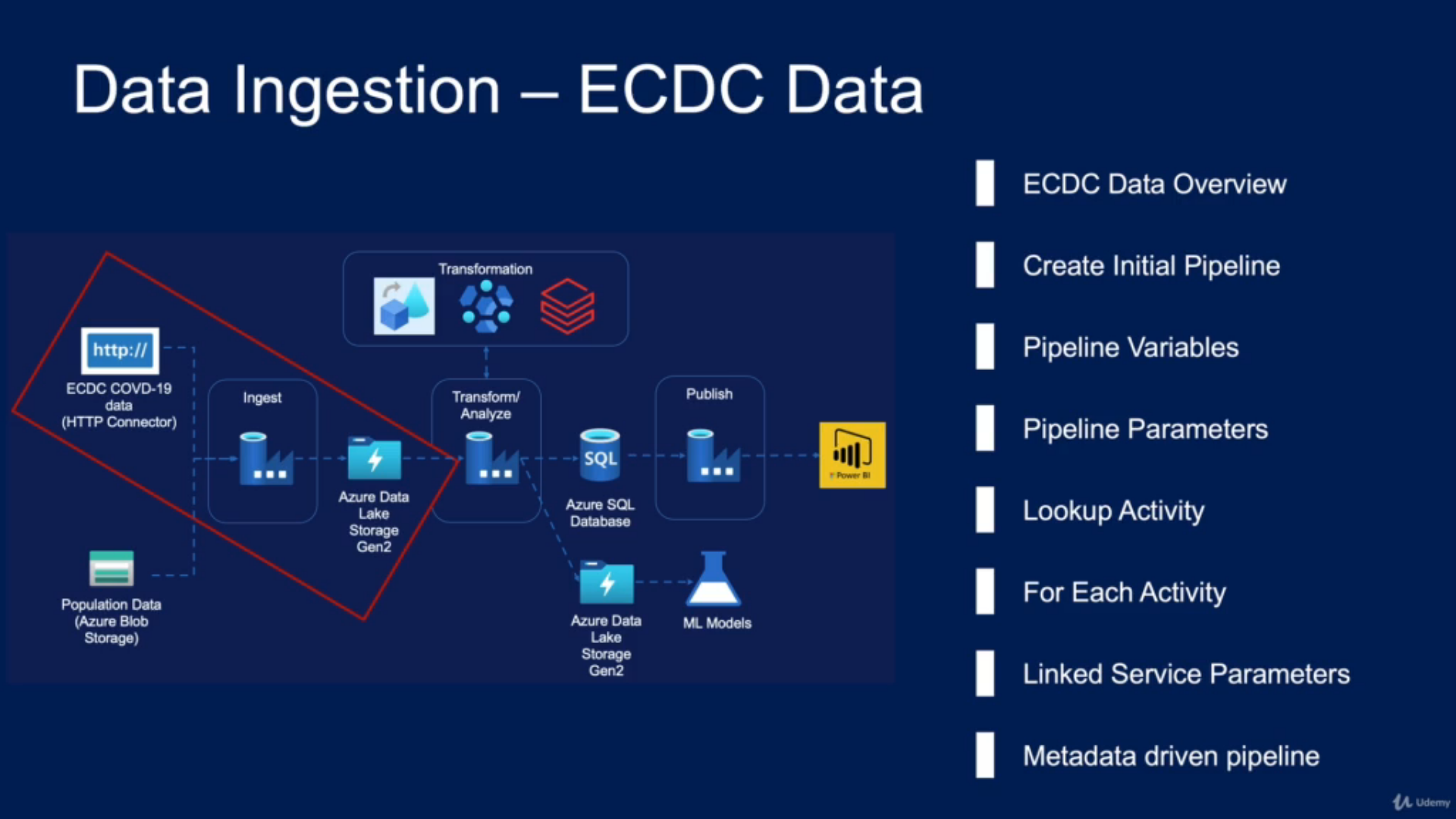




In order to use event trigger we have to use register Azure eventgrid resource Provider to our subscription.

Schedule triggers runs on specified time and date only.

Now , Data Ingestion from ECDC Website



Flow for Ingestion of Data:

1st create 2 storage account i.e 1) covidReportingds and 2) covidReportingAdf

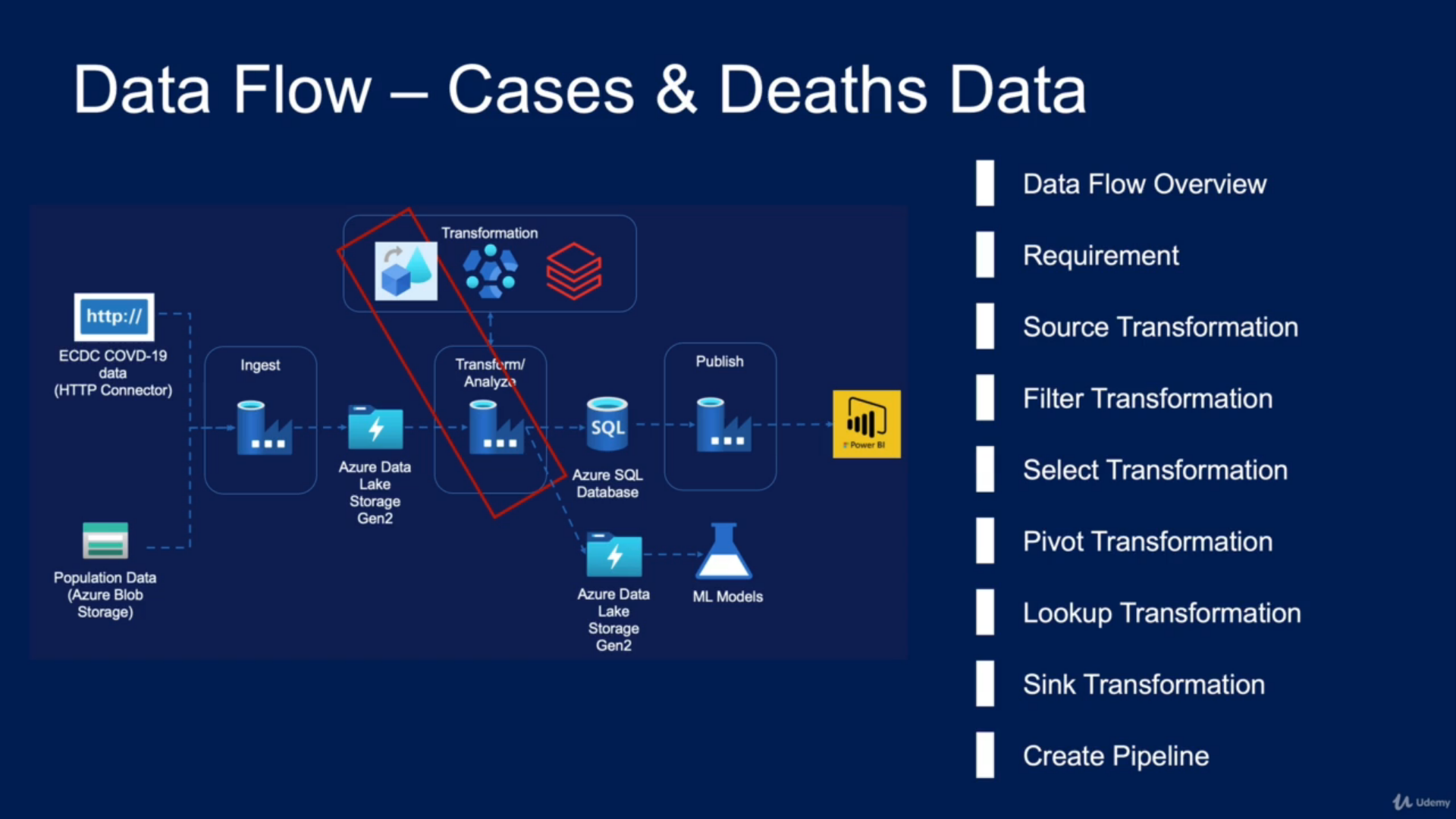
Create ADF Resource

Inside Adf Resource respective Source and Target Dataset -> then Create link service to link source file and Dataset inside Pipeline

Create Pipeline with Activity and connect dataset and complete the pipeline.

We can use same Pipeline for Different Source file having Same Activity And Operation with **Variables and Parameter**

Data Transformation





Select Transformation: This will help to select and remove specific columns and rename Columns.

HDinsigts:

This is not available for free subscription

HD insights allow access to database storage and Gen2lake using ManageIdentity Only.

Manage Identity give gives identity to various Azure resources.

What is Cluster?

* Cluster is the collection of computers/Nodes  that are networked together to perform some parallel computations on big data sets

In HDInsights we use HiveScript Code to Transform Data.