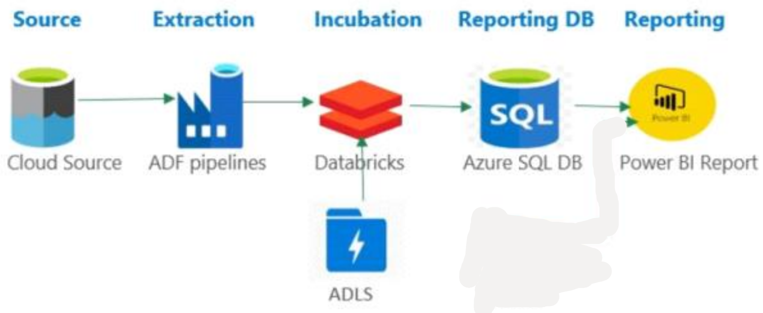
Requirements

30 July 2022 13:51



Metric needs to be shown

- Total number of flights by airline and airport on a monthly basis
- On time percentage of each airline for the year 2015
- Airlines with the largest number of delays
- Cancellation reasons by airport
- Delay reasons by airport
- Airline with the most unique routes

Airlines
Data Entity: Airlines
File Name: airlines.csv
Description: Airline codes and names

Airports
Data Entity: Airports
Location: airports.csv
Description: Airport codes, names, and locations

Flights
Data Entity: Flights
File Name: flights/partition-x.csv
Description: Flight Records

Find out the link of dataset here Dataset link

Source and Sink

30 July 2022 13:05

1. Setup free account on Azure Portal.
2. Create Blob Storage and create container in it(Give name : Source)
3. Upload given files into the source container(airlines.csv,airports.csv,flights.zip)
4. Create Blob Storage or ADLS GEN2(good to have) for target location(Give name: Sink)

Link for reference:

1. Setup free account in Azure Portal | Azure Data Factory Tutorial|
2. Setup Blob Storage

Azure Data Factory-- Sourcing

30 July 2022 13:19

1. Set up ADF for the project
2. Setup Azure KeyVault
3. Store required keys and secrets in the KeyVault(Grant ADF to read and list secrets and keys from KeyVault)
4. Create linked service and dataset to connect with source and sink blob storage
5. Setup Copy Activity to copy files from source container to sink container(Note: 3 file is zip make sure to unzip while copying from source)
6. Validate the count between source and sink using ADF

Link for reference:

1. Create and Setup KeyVault in Azure Data Factory
2. Put files into blob storage and create linked service and dataset
3. How to validate data between Source and Sink using Copy Activity in Azure Data Factory

AzureDataBricks

30 July 2022 13:32

1. Setup free community Edition account on Azure Databricks
2. Create two databases(raw and mart)
3. Connect Sink blob storage to load files into ADB
4. Create tables(delta preferred) or views or dataframe(preferred) and (if tables)store into raw database
5. Clean raw data, remove duplicates and do other required transformation.
6. Create fact and dimension table based on requirements and save it into mart database
7. After creating fact and dimension, publish that data into SQL DW(if possible) otherwise in SQL Database.

Link for reference:

1. Introduction to Databricks | How to setup Account |
2. How to read CSV file in PySpark
3. How to connect Blob Storage using SAS token using Databricks
4. How to write Dataframe with Partitions using PartitionBy in PySpark | Databricks Tutorial|
5. Difference between TempView and GlobalTempView
6. How to join two DataFrames in PySpark

Azure SQL/DW and POWERBI

30 July 2022 13:44

1. Create tables on Azure SQL DW/Database
2. Create Materialized views on top of tables
3. Download PowerBI Desktop
4. Connect Azure SQL with PowerBI
5. Import views and create relationship
6. Create Measures and report