**Data Ingestion End-to-End Pipeline with Azure**

Project 1

DESCRIPTION

Your company is looking for a data engineer and is inviting candidates to apply for this position by providing a portal where applicants can add their credentials.

As thousands of candidates have applied for this position, the company has a huge amount of data that it needs to upload to its website. This data is moved to Azure Data Lake Storage parallelly. The company wants to save the contents of all CSV files to Delta Lake of Azure Databricks so that these files can be retrieved and accessed from Azure Databricks when required.

**Steps to Perform:**

1. Create a landing storage account in Azure [DONE]

Graphical user interface, text, application, email

Description automatically generated

1. Store the CSV files in the storage account [DONE]

Graphical user interface, text, application, email

Description automatically generated

1. Create a staging storage account in Azure [DONE]

Graphical user interface, text, application, email

Description automatically generated

1. Create an Azure Data Factory resource and Azure Data Factory pipeline [DONE]

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

1. Create linked services for the storage accounts [DONE]

Graphical user interface, text, application, email

Description automatically generated

1. Use Azure Databricks as a part of the ADF pipeline [DONE]

Graphical user interface, application

Description automatically generated

1. Create a linked service in ADF for Databricks [DONE]

It was done but due to authorization block I couldn´t take a picture.

1. Convert the CSV files to Parquet files in staging storage [DONE]

Graphical user interface, application

Description automatically generated



1. Access Parquet files from the staging account in Azure Databricks [DONE]
2. Convert the Parquet files to Azure Databricks Delta tables [DONE]
3. Store and visualize the data from Azure Databricks Delta tables [DONE]

Steps 9, 10 and 11 were done but due to authorization block I couldn´t take pictures.

I am sending the code before the shutdown as well as the subsequent error messages.

On Databricks Notebook

import org.apache.spark.sql.functions.\_

dbutils.fs.mkdirs(“/FileStore/tables/input/”)

dbutils.fs.mkdirs(“/FileStore/tables/output/”)

input\_path = “/FileStore/tables/input/”

output\_path = “/FileStore/tables/output/”

storage\_account\_name = "stag1dataingestion"

storage\_account\_key = "ib879jaSKAN0ssad892"

container = "parquetFiles"

spark.conf.set(f"fs.azure.account.key.{storage\_account\_name}.blob.core.windows.net", storage\_account\_key)

dbutils.fs.ls(f"wasbs://{container}@{storage\_account\_name}.blob.core.windows.net/")

parquetDataFrame = spark.read.parquet(f"wasbs://{container}@{storage\_account\_name}.blob.core.windows.net/Candidate1.parquet")

parquetDataFrame.show()

parquetDataFrame.write.mode(“overwrite”).format(“parquet”).save(input\_path)

parquetDF = spark.read.parquet(input\_path)

parquetDF.show()

%sql

parquetDF.createOrReplaceTempView(“Candidate1”)

Select \* FROM Candidate1

Graphical user interface, text, application, website

Description automatically generated