# Green Health Data Lake POC

## ACRONYMS

ABS - Azure Blob Storage

ADF - Azure Data Factory

ADL - Azure Data Lake

SFTP - Secure File Transfer Protocol

ADB - Azure Databricks

DBFS - Databricks File System

AAD - Azure Active Directory

GH - Green Health

Contents

[Green Health Data Lake POC 1](#_Toc33729865)

[ACRONYMS 1](#_Toc33729866)

[INTRODUCTION 2](#_Toc33729867)

[POC OBJECTIVE 2](#_Toc33729868)

[GitLink 2](#_Toc33729869)

[STEP1 – COPY DATA FROM SFTP TO ABS USING ADF 2](#_Toc33729870)

[STEP2 – VALIDATE THE FILES IN ABS (PySpark) 2](#_Toc33729871)

[STEP3 – COPY DATA FROM ABS to ADL 3](#_Toc33729872)

[STEP4 – COPY DATE INTO ADB tables (SparkSQL) 3](#_Toc33729873)

[STEP5 – Loading to Azure SQLDataWarehouse 3](#_Toc33729874)

[STEP6 – Create Reports 4](#_Toc33729875)

## INTRODUCTION

Green Health is an Insurance company who sells health insurance products to customer for their health care services (hospitalization, walk-in, urgent care, emergencies, etc.). Different health care institutions and providers submit their claims to process for further calculation of payments and health factor of members (health factor helps Green Health to calculate their future premiums).

## POC OBJECTIVE

Green Health is creating centralized Data Lake to process, manage huge data in cost effective way.

1. Copy the raw inbound data from SFTP to ABS using ADF
2. Convert the files if not in readable format using ADB
3. Copy the files from ABS to ADL using ADF
4. Create Structured historical Source tables in ADB
5. Move Data to Azure SQL Datawarehouse using ADB or ADF
6. Create reports using Tableau/PowerBI

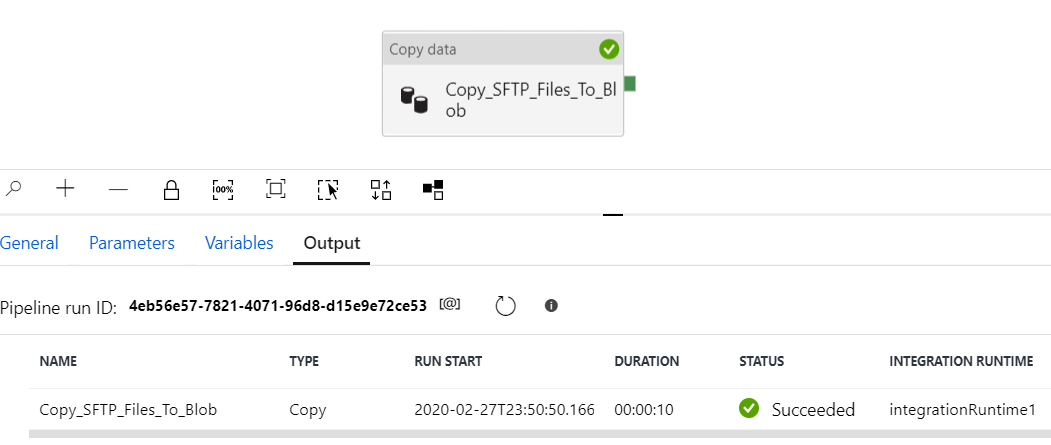
This POC is completed till step 4, it is under process to complete step 5 and 6

### GitLink

<https://github.com/vkantimahanti/Green-Health-Azure-Data-Lake>

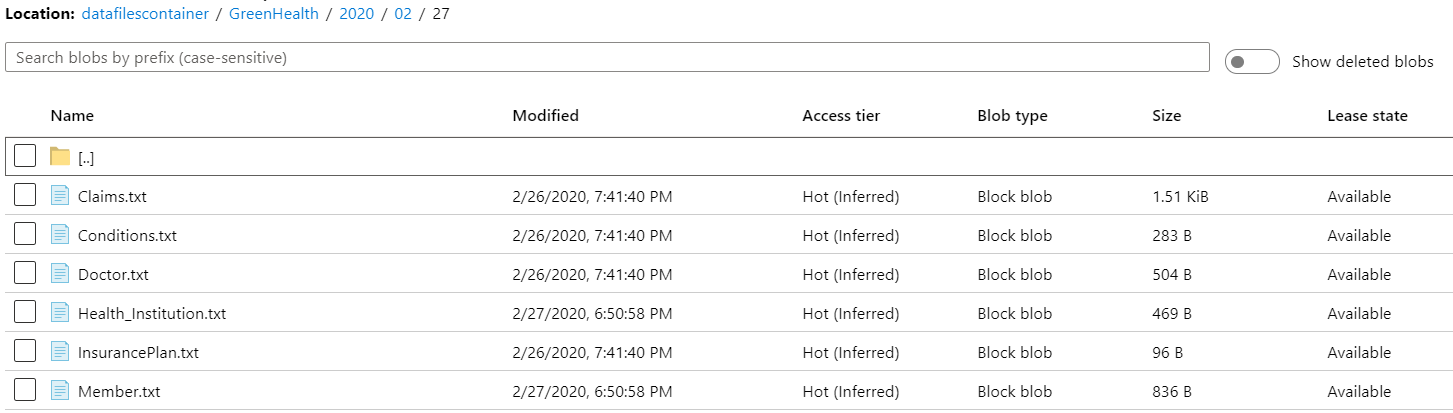
## STEP1 – COPY DATA FROM SFTP TO ABS USING ADF

* Created an ABS account with permanent account keys through Access keys.
* Installed integration run time in server and provided access to server machine in ADF.
* Created ADF pipeline with scripts to create dynamic folders while copying files from SFTP to Blob
* Source files in Git Link



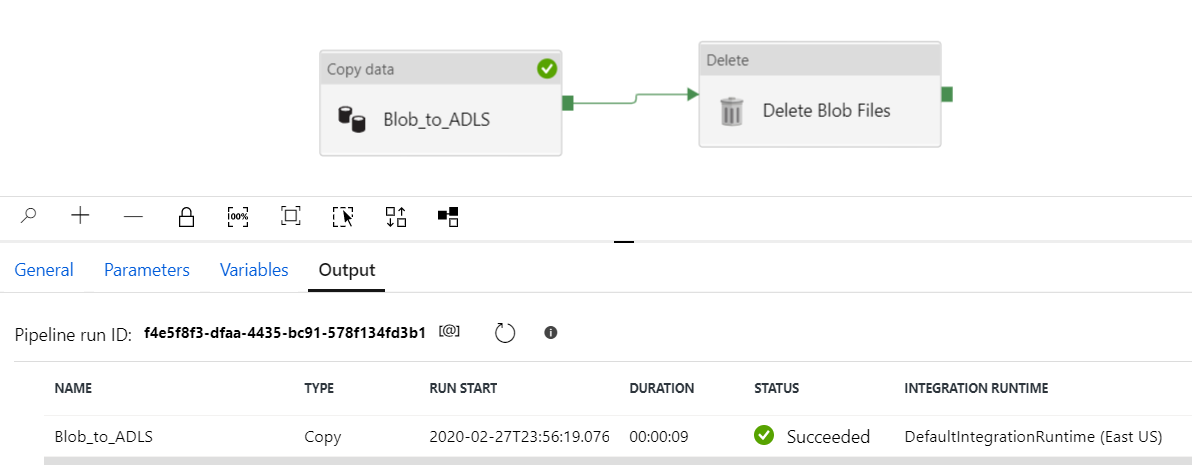
## STEP2 – VALIDATE THE FILES IN ABS (PySpark)

* Mounted Azure DBFS on Azure Blob using Account keys.
* Created PySpark code to convert non-readable files to readable format and perform validations
* If validation is successful, copy files from GH/YYYY/MM/DD to GH\_S else move to GH\_F.
* Drops the Azure DBFS mount (Mounting and validation code attached in Git link, ADBS.ipynb)



## STEP3 – COPY DATA FROM ABS to ADL

* Created ADL account with permanent Directory ID, Principal ID and ClientID using AAD.
* Created ADF pipeline to move data from ABS to ADL, also delete the files from ABS GH\_S.



## STEP4 – COPY DATE INTO ADB tables (SparkSQL)

* Create Temporary landing tables and load data from files in Azure DBFS using Spark SQL
* Create managed tables with Load\_DateTime (current date timestamp) as default values and load data from landing tables. These tables act as historical tables.
* Delete the temporary tables.
* Creating calculating views and loading back to tables (code attached in Git – DataLanding.sql)

## STEP5 – Loading to Azure SQLDataWarehouse

## STEP6 – Create Reports