SanKir Technologies https://www.sankir.com

email - info@sankir.com

POC on retail data using Databricks Spark Cluster, ADLS, Azure Key vaults

---

POC Objective

- Create Azure Databricks workspace and Spark Cluster
- Create SAS token using Shared access signature to access ADLS Gen2 Storage account
- Secure the sas token using Azure Key vaults
- Extract data from ADLS using ABFS connector
- Transform Data using Spark Transformations
- View the Data Profiling of retail data

About retail data set

This is a transactional data set which contains all the transactions for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data set path in S3

```
#Relative path of the data set (csv files) in ADLS with container and storgae
account under SanKir resourcegroup.
account_name = "retailsankir"
container_name = "data"
relative_path = "retail-data/q*"
```

Spark Configurations

```
# Spark configurations to access SAS token which is stored as secret in Key vault
spark.conf.set('fs.azure.account.auth.type.%s.dfs.core.windows.net' %
account_name, "SAS")
spark.conf.set('fs.azure.sas.token.provider.type.%s.dfs.core.windows.net' %
account_name, "org.apache.hadoop.fs.azurebfs.sas.FixedSASTokenProvider")
spark.conf.set('fs.azure.sas.fixed.token.%s.dfs.core.windows.net' % account_name,
dbutils.secrets.get(scope="sankir-scope", key="dbricks-abfs-sas"))
```

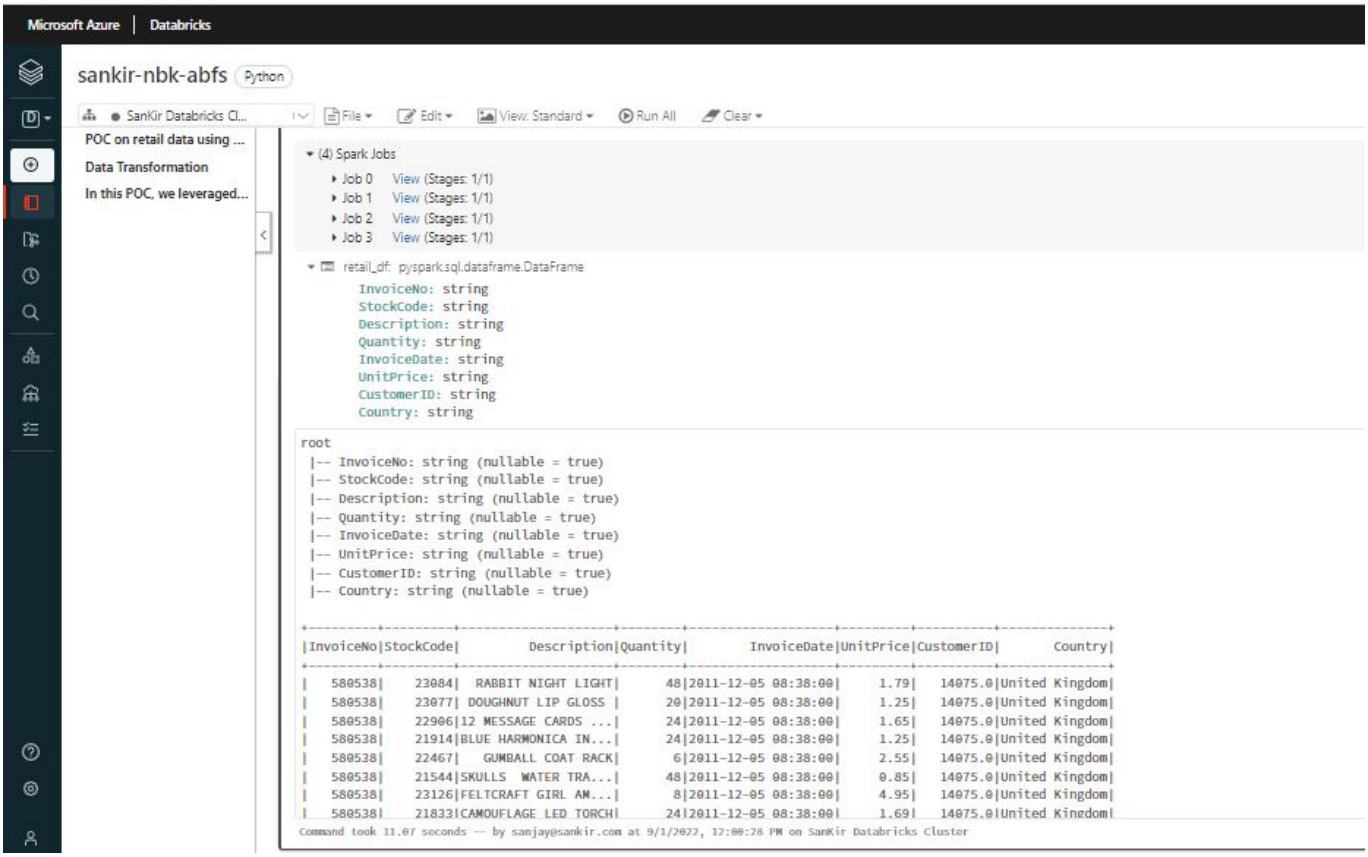
ABFS connector

```
# ABFS - Azure Blob File System
# ABFS Driver is used to access data in ADLS Gen 2 in a secured way
# ABFS Employs a URI format to address files and directories within ADLS Gen 2

# ABFS driver with DFS endpoint to access files in ADLS
# Create Saprk Dataframe by reading the csv files and show the Schema and content.

abfs_path = 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name,
account_name, relative_path)
retail_df = spark.read.option("header", "true").csv(abfs_path)
retail_df.printSchema()
retail_df.show()
```

Retail data - schema and records



Data Transformations using Spark

Use filter to show records with country name as poland.

```
# Data Transformation using Spark - filter records with country name as Poland
retail_df.filter(retail_df.Country == "Poland").show()
```

Cmd 7

```
1  # Data Transformation using Spark - filter records with country name as Poland
2  retail_df.filter(retail_df.Country == "Poland").show()
```

▸ (1) Spark Jobs

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
564819	23084	RABBIT NIGHT LIGHT	48	2011-08-30 12:11:00	1.79	12814.0	Poland
564819	POST	POSTAGE	1	2011-08-30 12:11:00	40.0	12814.0	Poland
574343	20839	FRENCH PAISLEY CU...	12	2011-11-04 10:11:00	0.85	12576.0	Poland
574343	20836	FRENCH PAISLEY CU...	12	2011-11-04 10:11:00	0.83	12576.0	Poland
574343	22722	SET OF 6 SPICE TI...	4	2011-11-04 10:11:00	3.95	12576.0	Poland
574343	21770	OPEN CLOSED METAL...	4	2011-11-04 10:11:00	4.95	12576.0	Poland
574343	22683	FRENCH BLUE METAL...	10	2011-11-04 10:11:00	1.25	12576.0	Poland
574343	84970L	SINGLE HEART ZINC...	12	2011-11-04 10:11:00	1.25	12576.0	Poland
574343	23144	ZINC T-LIGHT HOLD...	12	2011-11-04 10:11:00	0.83	12576.0	Poland
574343	22469	HEART OF WICKER S...	12	2011-11-04 10:11:00	1.65	12576.0	Poland
574343	82583	HOT BATHS METAL SIGN	12	2011-11-04 10:11:00	2.1	12576.0	Poland
574343	82580	BATHROOM METAL SIGN	12	2011-11-04 10:11:00	0.55	12576.0	Poland
574343	23250	VINTAGE RED TRIM ...	12	2011-11-04 10:11:00	1.25	12576.0	Poland
574343	23249	VINTAGE RED ENAME...	12	2011-11-04 10:11:00	1.65	12576.0	Poland
574343	85054	FRENCH ENAMEL POT...	6	2011-11-04 10:11:00	2.95	12576.0	Poland
574343	85053	FRENCH ENAMEL CAN...	6	2011-11-04 10:11:00	2.1	12576.0	Poland
574343	22340	NOEL GARLAND PAIN...	24	2011-11-04 10:11:00	0.39	12576.0	Poland
574343	35953	FOLKART STAR CHRI...	48	2011-11-04 10:11:00	0.39	12576.0	Poland

Command took 2.90 seconds -- by sanjay@sankir.com at 9/1/2022, 12:01:09 PM on SanKir Databricks Cluster

Use filter to show records with Quantity greater than 10.

```
# Data Transformation - filter - records with Qunatity greater than 10
qty10 = retail_df.filter(retail_df.Quantity > 10)
qty10.show()
```

Cmd 8

```
1  # Data Transformation - filter - records with Qunatity greater than 10
2  qty10 = retail_df.filter(retail_df.Quantity > 10)
3  qty10.show()
```

▸ (1) Spark Jobs

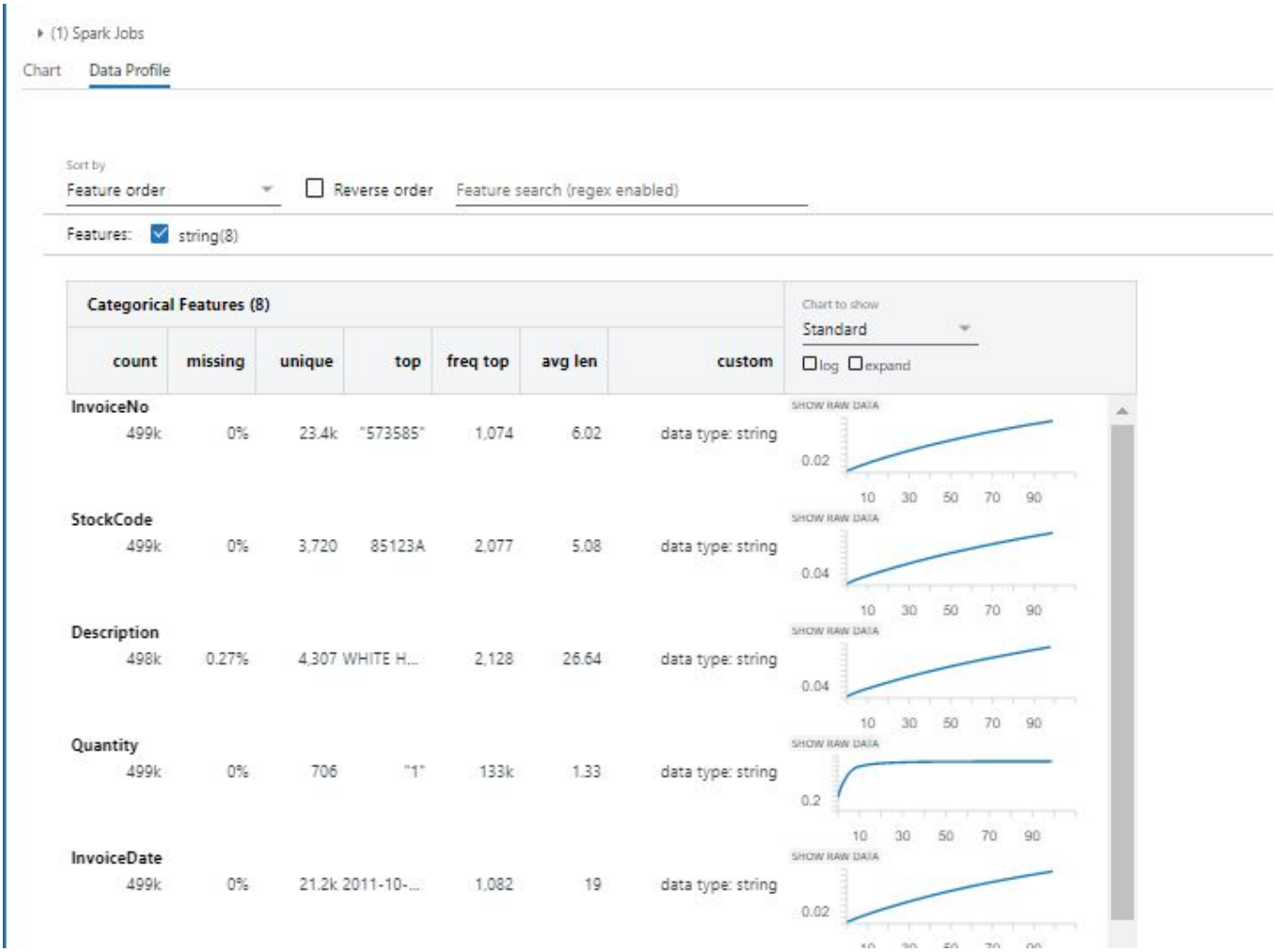
▸ qty10: pyspark.sql.dataframe.DataFrame = [InvoiceNo: string, StockCode: string ... 6 more fields]

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
580538	23084	RABBIT NIGHT LIGHT	48	2011-12-05 08:38:00	1.79	14075.0	United Kingdom
580538	23077	DOUGHNUT LIP GLOSS	20	2011-12-05 08:38:00	1.25	14075.0	United Kingdom
580538	22906	12 MESSAGE CARDS ...	24	2011-12-05 08:38:00	1.65	14075.0	United Kingdom
580538	21914	BLUE HARMONICA IN...	24	2011-12-05 08:38:00	1.25	14075.0	United Kingdom
580538	21544	SKULLS WATER TRA...	48	2011-12-05 08:38:00	0.85	14075.0	United Kingdom
580538	21833	CAMOUFLAGE LED TORCH	24	2011-12-05 08:38:00	1.69	14075.0	United Kingdom
580539	23235	STORAGE TIN VINTA...	12	2011-12-05 08:39:00	1.25	18180.0	United Kingdom
580539	22197	POPCORN HOLDER	36	2011-12-05 08:39:00	0.85	18180.0	United Kingdom
580539	22693	GROW A FLYTRAP OR...	24	2011-12-05 08:39:00	1.25	18180.0	United Kingdom
580539	22074	6 RIBBONS SHIMMER...	24	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22075	6 RIBBONS ELEGANT...	24	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22076	6 RIBBONS EMPIRE	24	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22389	PAPERWEIGHT SAVE ...	12	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22391	PAPERWEIGHT HOME ...	12	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22393	PAPERWEIGHT VINTA...	12	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22395	PAPERWEIGHT VINTA...	12	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	22481	BLACK TEA TOWEL C...	12	2011-12-05 08:39:00	0.39	18180.0	United Kingdom
580539	23320	GIANT 50'S CHRIST...	12	2011-12-05 08:39:00	1.25	18180.0	United Kingdom

Command took 0.35 seconds -- by sanjay@sankir.com at 9/1/2022, 12:01:20 PM on SanKir Databricks Cluster

Data profiling - is the process of examining, analyzing, and creating useful summaries of data

```
# Data profiling using display
display(retail_df)
```



Technologies leveraged in POC

- Databricks - 10.4 LTS (Spark 3.2.1)
- Databricks Notebook
- ADLS Gen2 Storage account
- Azure Key vaults
- Databricks secrets
- Apache Spark - pySpark 3.2.1