SanKir Technologies

Accelerating Data Engineering Solution on Cloud

Expertise in Data Engineering on Cloud

- Data Engineering pipeline – Architecture, Orchestration, Optimization and Monitoring
- End-to-End Automation
  - AWS CDK Toolkit
  - Terraform
- Databricks, Apache Airflow
- Big Data - Apache Spark, Hadoop, HDFS
- DWH - Google BigQuery, Snowflake and DBT

aws Google Cloud Azure

Consultancy

Engagement Partner for Business Opportunities

consultancy.sankir.com

CONSULTANCY

Solution Architecture

- Architecting the Data Pipeline on Cloud
  - Multiple Solution options for a data problem
- End-to-End DE Automation using AWS CDK Toolkit and Terraform
- Data Engineering Solution on Cloud
  - Optimizing the Data processing
  - Best Practices in data & network security
  - Spark Cluster - Sizing and Optimization for better performance of workloads
- Leverage SanKir AWS/Azure/GCP Cloud Infrastructure to quickly test the solution
- SanKir can Work with Client’s CxO, Technical Managers or Engineering team to solve the organizational Data problems
- SanKir aims to align with Client’s Core Business Objectives

Leverage SanKir services in

- Cloud Services and Solution
- Data Engineering Tools and Solution
- CI/CD – DevOps
- Container Orchestration using Docker and Kubernetes
- Streaming using Spark streaming and Kafka
- API development using Golang

PoC on Data Engineering pipeline powered by SanKir Framework and Assets

Consultancy in following Cloud Services and Solution

AWS

- Amazon S3
- Spark Cluster – AWS EMR
  - Cluster Sizing and Monitoring
- Cloud DWH - RedShift/Athena
- AWS CDK Toolkit
- AWS SDK using Boto3
- Orchestration
  - AWS Cloud Formation
  - AWS Lambda
  - CloudWatch
- IAM and VPC
- Secrets Manager, KMS
- RDS - PostgreSQL, MySQL
- Apache Airflow/MWAA
- Databricks on AWS
- Snowflake on AWS
- CodeBuild, CodeDeploy
- AWS Cost Management

Azure

- ADLS
- Spark Cluster - Azure Databricks
- Cloud DWH
  - Azure Synapse Analytics
  - SQL Databases
- ABFS(S)/WASB(S) - API to access ADLS
- IAM
- Access Keys
- Shared Access Signature
- Azure Key Vaults, Keys & Secrets
- Service Principal
- Azure Data Factory
- Snowflake on Azure
- Azure Active Directory (AAD)

GCP

- Google Cloud Storage
- Spark Cluster - Google Dataproc
- Cloud DWH - Google BigQuery
- Google Cloud SDK
  - Google Cloud CLI
  - Cloud Shell
- IAM
- Cloud Functions
- Secret Manager
- Cloud Run
- Cloud Build
- Terraform on GCP

Consultancy in following Data Engineering Tools and Solution

Databricks

- ELT Solution provided on Azure, Azure and GCP

- Databricks Spark Cluster
    - Cluster Types - All Purpose and Job Clusters
    - Cluster Configuration
    - Cluster Mode - Standard, High Concurrency, Single Node
    - Cluster Pools

- Databricks Runtime
- Auto Scaling & Auto Termination

- Notebook Workflow Utilities

- DBFS & Databricks mounts
- Databricks secrets
- File System Utilities

Snowflake

- ELT Solution provided on AWS, Azure and GCP

- Loading Data from S3/ADLS/GCS
- DML for bulk data loading/unloading using COPY command
- Snow pipe - Load data fast, analyze even faster

- Snowflake Connector for Python & Spark
- Accelerating BI Queries with Caching

- Eliminating Concurrency Issues with Snowflake Virtual Warehouses
- Data Protection with Time Travel
- Zero copy cloning

- Standard and Extended SQL
- Advanced DML such as multi-table INSERT, MERGE, and multi-merge

- Statistical and Analytical aggregate functions
- Windowing functions

Apache Airflow

- Data Pipeline Solution on AWS, Azure and GCP
- Dynamic DAG Authoring using task groups
- DAG Dependencies

- Airflow Variables
- Scheduling and Triggers

- Templating
- Task Flow API
- XCOMs

Data Build Tool - DBT

- Types of Materializations, Incremental Models & Ephermal Models

- Hooks, Snapshots & Macros

- Integrate DBT with Snowflake, Databricks, BigQuery, RedShift & Apache Spark

SanKir as Engagement Partner for Business Opportunities

- Architecting the Solution for Client’s Potential Business opportunities
  - Recommendation on right choice of Cloud vendor, Data Pipeline Tools and services and Cloud Data Warehouse based on Client requirements
- Execute PoC to win a Project deal
- Customized technology upskill program for Corporate professionals
  - Enable Client to showcase their team’s skills for resourcing opportunities
  - Hands-on sessions with PoC project
- Joint Go-To-Market Strategy
  - SanKir can partner with Client from initial bid stage to Architecting the Solution to Project Execution

WHY SANKIR

- Differentiator in Pricing
  - Competitive pricing compared to Cloud vendor's professional services
  - Aim to deliver within Client's Budget

- Industry Experience in Multitude of Data Projects

Let SanKir own your Data Problems

Proof of Work - PoC

- PoCs done for Clients
  - Retail data transformation using Spark, S3, AWS EMR, Athena
    - End-to-End DE Automation using AWS CDK Toolkit
      - EC2 creation, Airflow installation, EMR creation and Spark job submission
    - Data Profiling – Column profile detail & Data quality metrics
  - Orchestration using Airflow for Ed-Fi operational store used in K-12 Education
    - Dynamic DAGs
    - Schedule and monitoring of task-groups and tasks
  - Data Engineering using built-in and scalable Azure Databricks platform
    - Infra creation (Spark Cluster) – compute sizing
    - Data Transformation, Best practice using Key vaults
  - Loading retail data from AWS S3 file storage into Snowflake tables
  - Designing Data pipeline for ETH blockchain using Airflow and AWS EMR

Cost of Cloud Infrastructure of PoC for the Organization will be very minimal

PoC – E2E Automated DE Pipeline using Airflow and AWS EMR for Retail data set

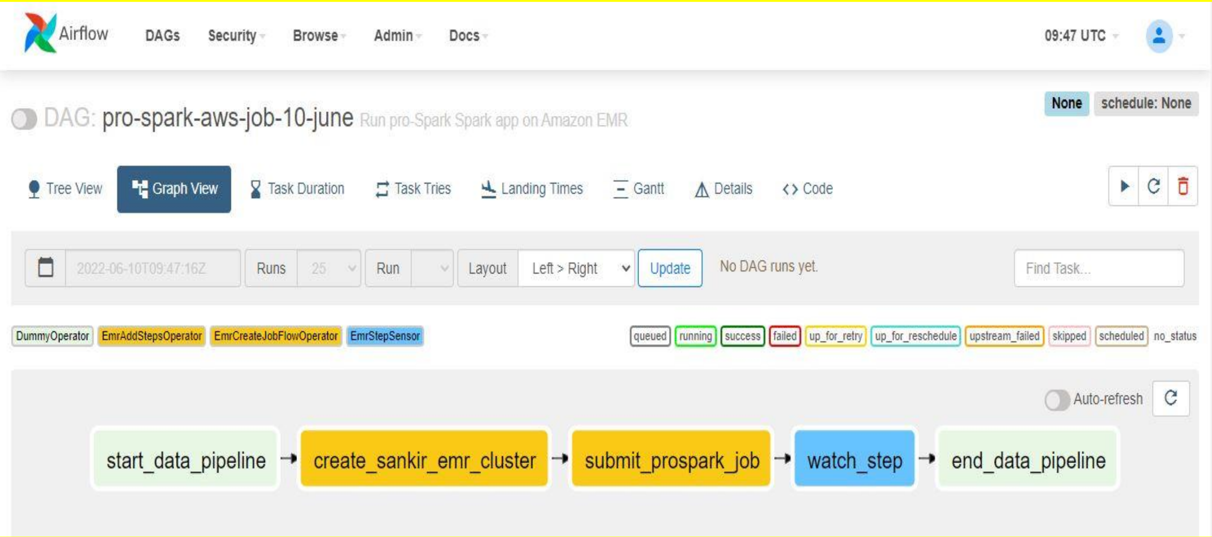
Data Size	1 GB
AWS Services	S3, Athena, EMR

AWS EMR – Spark Cluster	
Nodes	1 Master and 2 Cores (3-Node Cluster)
Hardware	Instance type - m5.4xlarge 16 vCore, 64 GiB memory, EBS only storage EBS Storage:256 GiB
Release label	emr-5.35.0
Hadoop distribution	Hadoop 2.10.1
Applications	Spark 2.4.8 ; Ganglia 3.7.2 ; Airflow(MWAA) 2.0.2

- Spark Cluster sizing based on Hardware Configuration
- Spark Job monitoring using Spark UI/Ganglia
- Cost of Cloud Infrastructure of PoC for the Organization will be very minimal
- PoC setup can be customized to customer needs within a week

Full details of PoC will be shared upon request

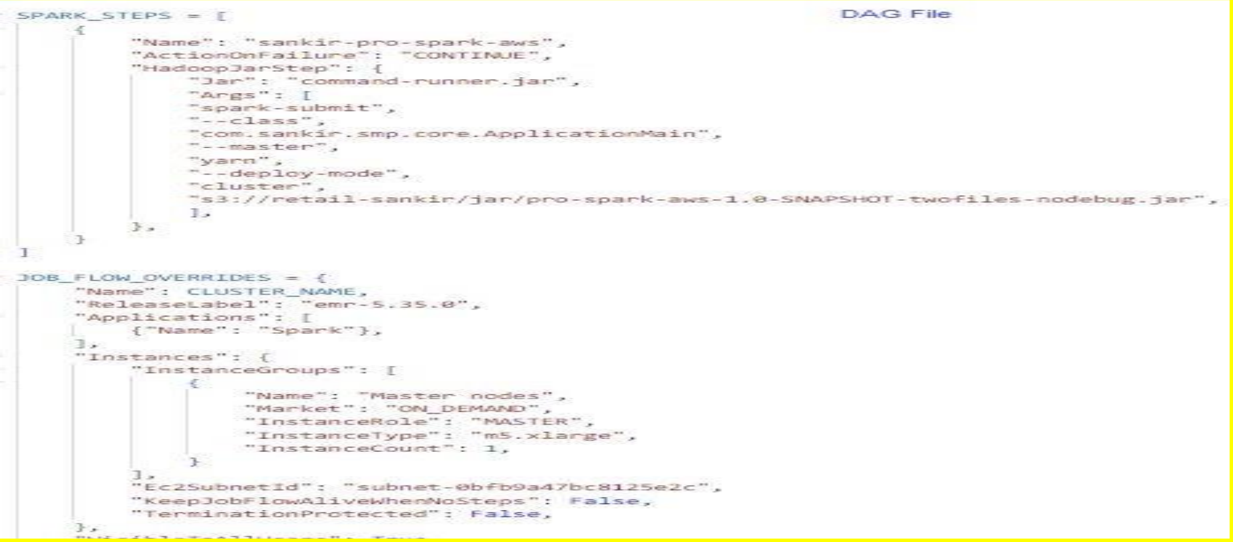
E2E DE Pipeline for Retail data set - Screenshots



DAG: pro-spark-aws-job-10-june Run pro-Spark Spark app on Amazon EMR

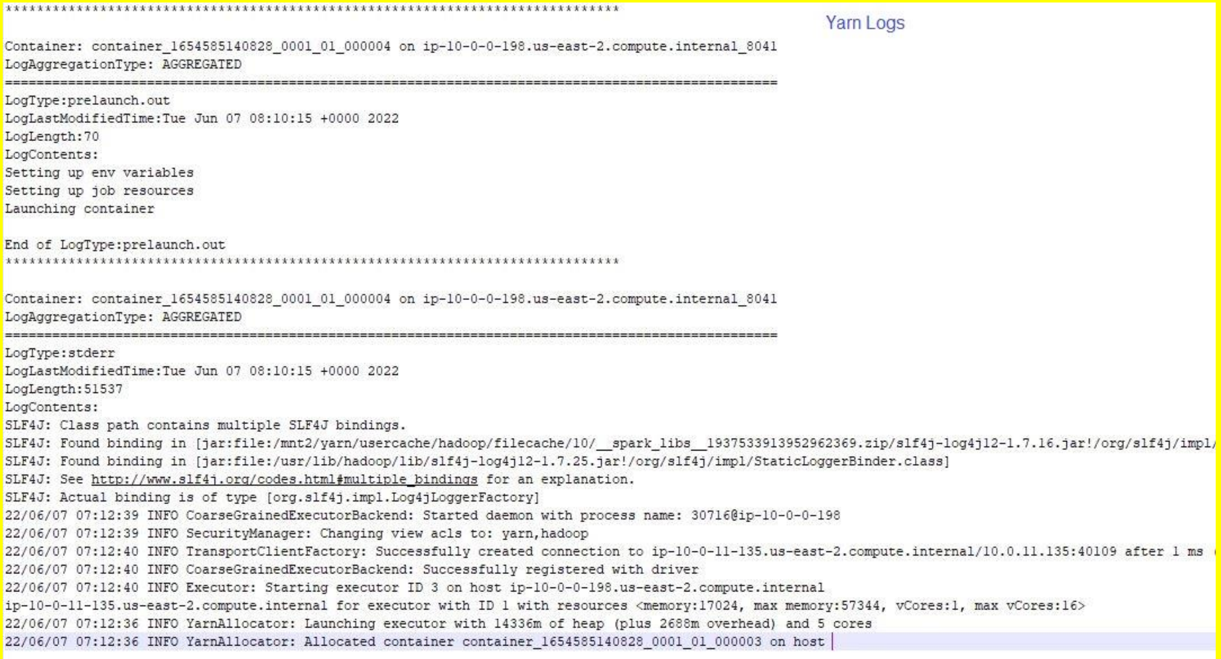
None schedule: None

2022-06-10T09:47:16Z Runs 25 Run Layout Left > Right Update No DAG runs yet. Find Task...



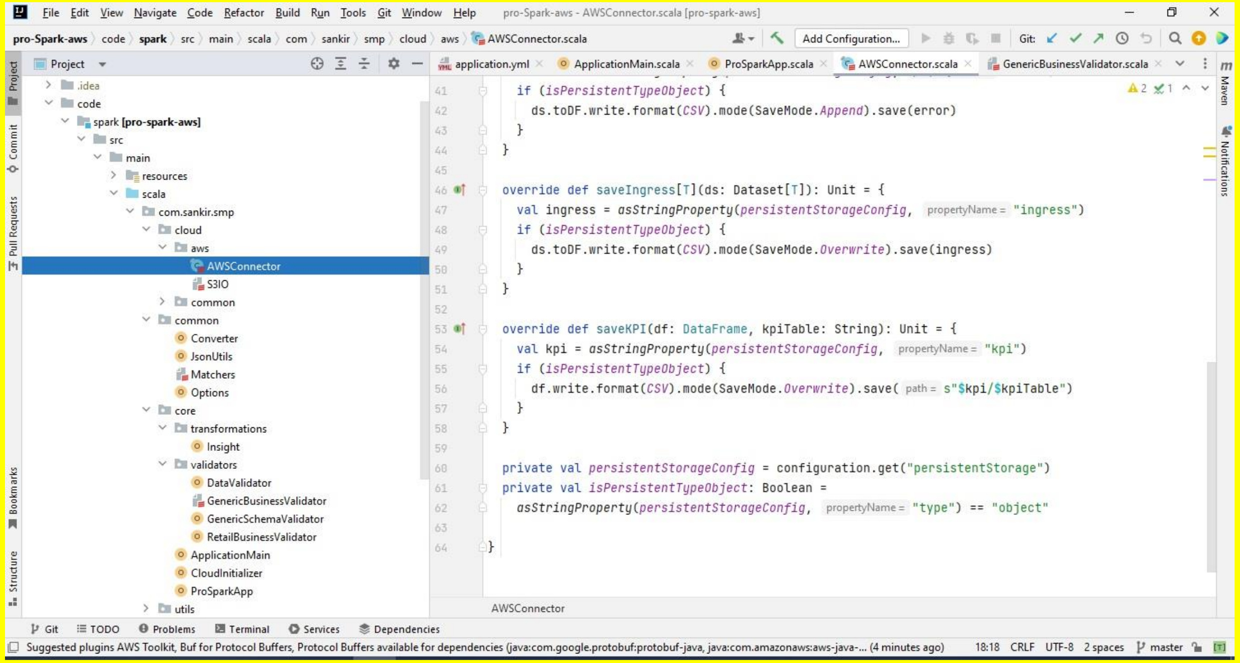
DAG File

```
SPARK_STEPS = [
    {
        "Name": "sankir-pro-spark-aws",
        "ActionOnFailure": "CONTINUE",
        "HadoopJarStep": {
            "Jar": "command-runner.jar",
            "Args": [
            "spark-submit",
            "--class",
            "com.sankir.smp.core.ApplicationMain",
            "--master",
            "yarn",
            "--deploy-mode",
            "cluster",
            "s3://retail-sankir/jar/pro-spark-aws-1.0-SNAPSHOT-twofiles-nodebug.jar",
            ],
        },
    }
]

JOB_FLOW_OVERRIDES = {
    "Name": CLUSTER_NAME,
    "ReleaseLabel": "emr-5.35.0",
    "Applications": [
        {"Name": "Spark"},
    ],
    "Instances": {
        "InstanceGroups": [
            {
                "Name": "Master nodes",
                "Market": "ON_DEMAND",
                "InstanceRole": "MASTER",
                "InstanceType": "m5.xlarge",
                "InstanceCount": 1,
            }
        ],
        "Ec2SubnetId": "subnet-0bfb9a47bc8125e2c",
        "KeepJobFlowAliveWhenNoSteps": False,
        "TerminationProtected": False,
    },
```



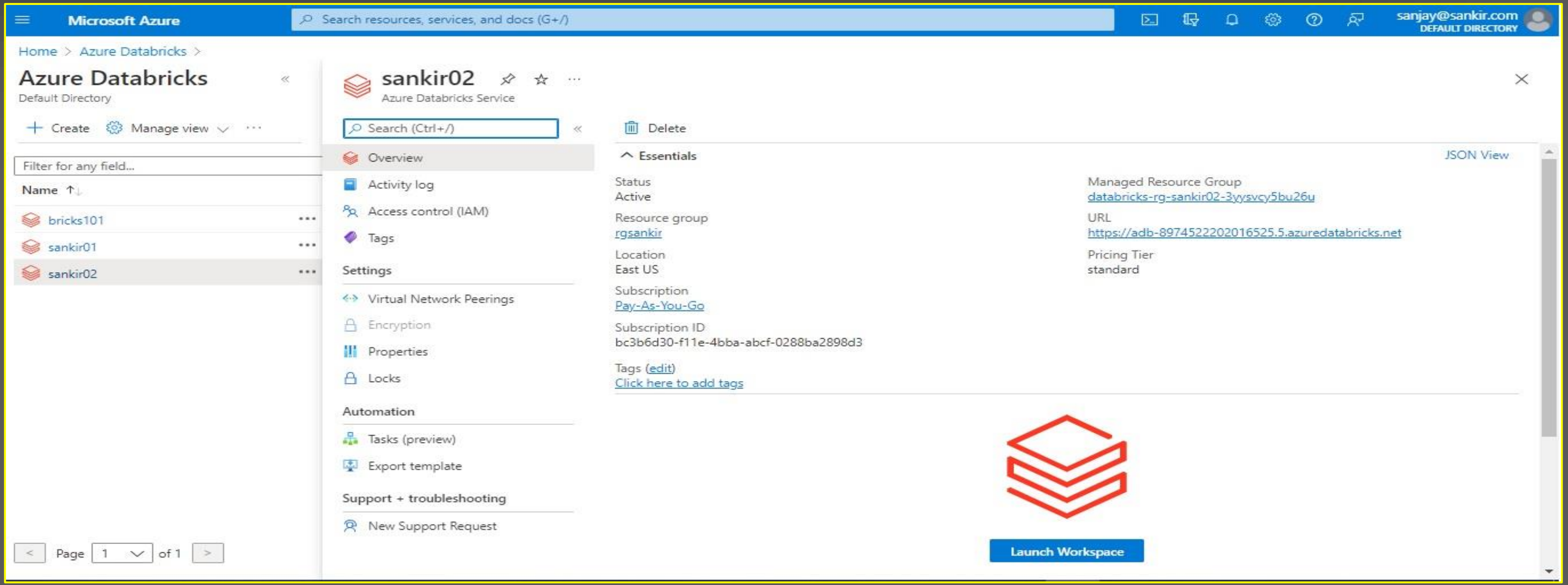
Yarn Logs

```
***********************************************************************************

Container: container_1654585140828_0001_01_000004 on ip-10-0-0-198.us-east-2.compute.internal_8041
LogAggregationType: AGGREGATED
====================================================================================================
LogType:prelaunch.out
LogLastModifiedTime:Tue Jun 07 08:10:15 +0000 2022
LogLength:70
LogContents:
Setting up env variables
Setting up job resources
Launching container

End of LogType:prelaunch.out
***********************************************************************************

Container: container_1654585140828_0001_01_000004 on ip-10-0-0-198.us-east-2.compute.internal_8041
LogAggregationType: AGGREGATED
====================================================================================================
LogType:stderr
LogLastModifiedTime:Tue Jun 07 08:10:15 +0000 2022
LogLength:51537
LogContents:
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/mnt2/yarn/usercache/hadoop/filecache/10/__spark_libs__1937533913952962369.zip/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/06/07 07:12:39 INFO CoarseGrainedExecutorBackend: Started daemon with process name: 30716@ip-10-0-0-198
22/06/07 07:12:39 INFO SecurityManager: Changing view acls to: yarn,hadoop
22/06/07 07:12:40 INFO TransportClientFactory: Successfully created connection to ip-10-0-11-135.us-east-2.compute.internal/10.0.11.135:40109 after 1 ms
22/06/07 07:12:40 INFO CoarseGrainedExecutorBackend: Successfully registered with driver
22/06/07 07:12:40 INFO Executor: Starting executor ID 3 on host ip-10-0-0-198.us-east-2.compute.internal
ip-10-0-11-135.us-east-2.compute.internal for executor with ID 1 with resources <memory:17024, max memory:57344, vCores:1, max vCores:16>
22/06/07 07:12:36 INFO YarnAllocator: Launching executor with 14336m of heap (plus 2688m overhead) and 5 cores
22/06/07 07:12:36 INFO YarnAllocator: Allocated container container_1654585140828_0001_01_000003 on host
```


File Edit View Navigate Code Refactor Build Run Tools Git Window Help
pro-Spark-aws - AWSConnector.scala [pro-spark-aws]
pro-Spark-aws › code › spark › src › main › scala › com › sankir › smp › cloud › aws › AWSConnector.scala
Add Configuration...
Project
application.yml
ApplicationMain.scala
ProSparkApp.scala
AWSConnector.scala
GenericBusinessValidator.scala
.idea
code
spark [pro-spark-aws]
src
main
resources
scala
com.sankir.smp
cloud
aws
AWSConnector
S3IO
common
common
Converter
JsonUtils
Matchers
Options
core
transformations
Insight
validators
DataValidator
GenericBusinessValidator
GenericSchemaValidator
RetailBusinessValidator
ApplicationMain
CloudInitializer
ProSparkApp
utils
AWSConnector
Git
TODO
Problems
Terminal
Services
Dependencies
Suggested plugins AWS Toolkit, Buf for Protocol Buffers, Protocol Buffers available for dependencies (java:com.google.protobuf:protobuf-java, java:com.amazonaws:aws-java-... (4 minutes ago)
18:18 CRLF UTF-8 2 spaces master

```
    if (isPersistentTypeObject) {
      ds.toDF.write.format(CSV).mode(SaveMode.Append).save(error)
    }
  }

  override def saveIngress[T](ds: Dataset[T]): Unit = {
    val ingress = asStringProperty(persistentStorageConfig, propertyName = "ingress")
    if (isPersistentTypeObject) {
      ds.toDF.write.format(CSV).mode(SaveMode.Overwrite).save(ingress)
    }
  }

  override def saveKPI(df: DataFrame, kpiTable: String): Unit = {
    val kpi = asStringProperty(persistentStorageConfig, propertyName = "kpi")
    if (isPersistentTypeObject) {
      df.write.format(CSV).mode(SaveMode.Overwrite).save( path = s"$kpi/$kpiTable")
    }
  }

  private val persistentStorageConfig = configuration.get("persistentStorage")
  private val isPersistentTypeObject: Boolean =
    asStringProperty(persistentStorageConfig, propertyName = "type") == "object"

}
```

PoC - Data Engineering using built-in and scalable Azure Databricks platform



Status: Active

Resource group: rgsankir

Location: East US

Subscription: Pay-As-You-Go

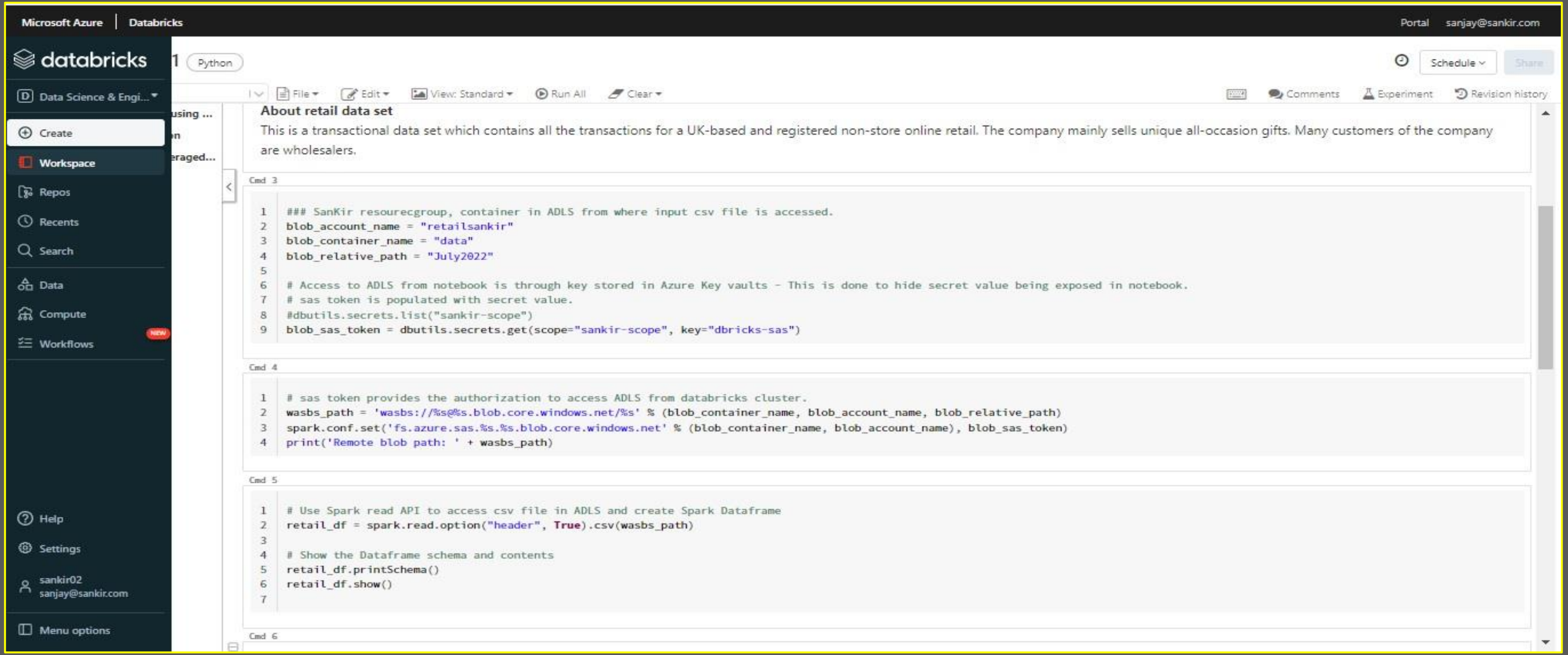
Subscription ID: bc3b6d30-f11e-4bba-abcf-0288ba2898d3

Tags (edit): Click here to add tags

Managed Resource Group: databricks-rg-sankir02-3yysvcy5bu26u

URL: https://adb-8974522202016525.5.azuredatabricks.net

Pricing Tier: standard



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.

Cmd 3

```
### SanKir resourcegroup, container in ADLS from where input csv file is accessed.
blob_account_name = "retailsankir"
blob_container_name = "data"
blob_relative_path = "July2022"

# Access to ADLS from notebook is through key stored in Azure Key vaults - This is done to hide secret value being exposed in notebook.
# sas token is populated with secret value.
#dbutils.secrets.list("sankir-scope")
blob_sas_token = dbutils.secrets.get(scope="sankir-scope", key="dbricks-sas")
```

Cmd 4

```
# sas token provides the authorization to access ADLS from databricks cluster.
wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blob_account_name, blob_relative_path)
spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name, blob_account_name), blob_sas_token)
print('Remote blob path: ' + wasbs_path)
```

Cmd 5

```
# Use Spark read API to access csv file in ADLS and create Spark Dataframe
retail_df = spark.read.option("header", True).csv(wasbs_path)

# Show the Dataframe schema and contents
retail_df.printSchema()
retail_df.show()

```

Cmd 6

SanKir Team Profile

Kiran Hiremath, Director

IT Professional with 27+ years of Experience.

- Expert in Data Engineering, Cloud Services and Distributed Computing using Apache Spark
  - Data Engineering Pipeline – Architecture, Orchestration, Optimization and Monitoring
  - Spark Cluster sizing – AWS EMR, Google Cloud Dataproc
  - Cloud Storage and Datawarehouse – AWS S3, Athena, ADLS, GCS, BigQuery
  - Big Data Technologies: Hadoop, HDFS, Spark, Scala, pySpark
  - Databricks on Azure & AWS
- Experience in Pre-Sales, CoE, Alliance, Software Development & Management
- Worked for TCS, Wipro and has interfaced with MNCs like Nortel, Motorola and Alcatel-Lucent
  - Feature design in Network-Switches
  - Data-driven Contact Centre solution in Healthcare, Banking and Telcos
  - Data Migration, ETL to Salesforce.com application from Siebel CRM
  - IT Transformation Programs

GitHub https://github.com/kiranhm1972

in https://www.linkedin.com/in/hiremath-kiran/

Experience in Consultancy, Conceptualization, Asset & Solution Development on Data Platforms

SanKir Team Profile

Sanjay Bheemasenrao, Director

27+ years of experience in building products and services

- Demonstrated leadership in building focused teams to achieve excellence
- Functioned in various capacities and has held many important positions in India and US. Has worked for reputed companies like Oracle, Tata Elxsi, GE and Prentice Hall.
- Expert in Data Engineering covering all the Data Management aspects
  - Data Engineering Pipeline – Architecture, Orchestration, Optimization and Monitoring
  - Big Data Technologies: Hadoop, HDFS, Spark, Scala, Python, Hive
  - Architecting Distributed Computing solutions using Apache Spark, Hive, BigQuery and AWS Athena, RedShift, DBT, AWS Lambda, Cloud Watch, Cloud Build, Apache Airflow.
- Worked in Oracle India for 16+ years
  - Oracle E-Business Release Management expert - 11i to R12.2
  - Developed SaaS solutions for Manufacturing industry on Oracle Cloud

GitHub https://github.com/sbheemas

in https://www.linkedin.com/in/sanjaybheemasenarao/

Technology focus areas – AWS and GCP, Data Lake and Cloud Data warehouses

Additional Capabilities

Self-paced Online Courses

* Please visit www.sankir.com to download Detailed Brochure of Self-paced online courses

For Experienced IT professionals

Data Engineering on Cloud

aws

pro-Spark-aws

Google Cloud

pro-Spark-gcp

Snowflake with pro-Spark

snowflake

pro-Spark on aws

snowflake

pro-Spark on gcp

Databricks on Cloud

Learn Databricks on AWS

Learn Databricks on GCP

- Quiz
- Assignment
- Bonus Lessons

Exclusive batches for corporate at competitive price

About SanKir

10,000+

Lines Of Code

1000+

Coached Professionals

20+

Meetups

55+

IT Industry Experience

SanKir History

2022 – Consultancy

Solution Architecture, PoC, Cloud Services & Solution, Data Engineering Tools & Solution

2020 – Self-paced Online Courses

Data Engineering on Cloud – AWS, Azure and GCP.

2018 - Founded

Classroom Courses on Big Data, Spark, Java with Capstone Projects.

Thank you!

🌐 consultancy.sankir.com

✉ info@sankir.com

in

YouTube