BCG CASE STUDY Notes

I have executed various scenarios on Databricks. Below is the access link to notebook

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3435925290439665/3209428621457615/6193880863326018/latest.html>

Also I have shared a zip file named “HAASBCG” which resonates with the best methods and practises followed in the Cloudera Hue environement.

Below is the dummy folder structure in HDFS

Graphical user interface, application

Description automatically generated

1. All the dependencies jars and files required by spark-submit will be stored inside the Dependencies and will be fetched by edge node when the job is triggered.
2. All incoming files will be placed in the inbound folder which we can further sub divide to landing, staging and archive depending on the business needs and transformation logics.
3. The spark scala jar file will be placed in the Jar folder from where it will be fetched by edge node before spark submit is triggered.
4. The output of the spark job will be stored in outbound. The resultant output of each analytics scenario will be placed under Analytics\* folders.
5. The Shell script and the coordinator ,workflow will be stored inside Shell and Workflows respectively
6. The job.properties file will be used trigger the coordinator.