**Name: Abhishek Obla Hema**

**Email:** [**oh\_abhishek@yahoo.com**](mailto:oh_abhishek@yahoo.com)

**Spark Assignment-1(Marketing Campaign Analysis) Documentation**

This file details and describes all the attached files for the Spark Assignment- 1

**Tools Used:**

1. Python3 – Microsoft VScode
2. Apache Spark (On GCP Cluster)
3. GCP services – DataProc/GCS
4. Jupyter Lab
5. Apache Hive

**Files Attached:**

1. Spark\_Ass1.pdf – This file
2. Spark\_Marketingdata– Pyspark File that details all the analysis
3. HQL\_datastore.txt – HQL commands used to create external tables off the data

**Process and File Descriptions:**

**Step 1:**

First I created three dataframes:

* Campaigns
* Users
* Stores

A screenshot of a computer

Description automatically generated

**Step 2:**

I then performed the necessary transformations that has been asked in the question and placed the output in separate directories in the HDFS .



A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Step 3:**

We can check the output has been saved using the UI

A screenshot of a computer

Description automatically generated

**Step 4:**

Using HQL statements I then created Hive external tables using JSON serde.

A screenshot of a computer program

Description automatically generated

**Step 5:**

Querying Hive tables to check the output data to make sure

A screenshot of a computer

Description automatically generated