**ASSIGNMENT SOLUTIONS**

1. Please prepare a proposal (any format is fine) of how you would tackle this project, what tools/programming language would you use, how would you model this type of data set, where would you do this modelling?

**Answer**:

Tools/languages need to use.

* Python -- Programming language
* Apache Spark -- Distributed processing system
* Snowflake -- database, better in handling huge volume of data.
* Apache Airflow -- Scheduling and Monitoring
* AWS EMR cluster -- for running spark code, preferably to use eu-west-2 region in London
* Git and Jenkins -- Deployment and version management

PFA the data flow diagram and schema.

 

1. Using Python programming language, please prepare a script to ingest the provided csv file, perform some sample manipulation that you feel relevant to the data set, and load the data to a Microsoft or Snowflake database of your choosing.

**Answer**:

PFA the batch process code.



1. Load the data into a Microsoft or Snowflake SQL database, and then prepare SQL to perform some manipulation to the data set that you feel appropriate to do some sample modelling of the data.

**Answer**:

PFA the SQL queries used.

