**Data Engineering / CDW\_SAPP Credit Card System – project outline**

**\*\*\* please follow instructions in the GitHub repository’s README.md file \*\*\***

**Core Java**

* The application aggregates nearly 50K rows of transaction/bank/customer data for optimal Analysis and BI.
* Inside the ‘**src**’ folder: 1 – ‘**models**’ folder contains Transaction & Customer model classes that are used to instantiate those objects. 2 – ‘**drivers**’ folder contains files with logic that runs the program. 3 – ‘**data\_access\_objects**’ folder contains DAO wrapper classes that fetch data from the RDBMS using JDBC. 4 – ‘**resources**’ folder contains a *SQL queries* file that is accessed by DAO objects, and a *User Input Validation Methods* file that is accessed by Driver classes.

**RDBMS/mySQL Description**:

* The file CDW\_SAPP.sql is the original database, the file CDW\_SAPP\_FINAL.sql is the final database.
* **/ETL/ Final\_Database\_Export/**

**Hadoop/hdfs/dataware housing**

* Data is extracted to staging tables in HDFS (located at /user/maria\_dev/Credit\_Card\_System/).
* **/ETL/ ETL\_req\_2-2-1\_Sqoop/**

**Hive and Partition**

* Data is loaded into Hive Warehouse as ORC files (/apps/hive/warehouse) in a partitioned fashion where applicable.
* Credit Card table is partitioned by transaction type, Branch table is partitioned by branch state, Time table is portioned by quarter.
* Notice that Customer table is not partitioned.
* **/ETL/ ETL\_req\_2-2-2\_Hive/**

**Oozie (Sqoop and Hive)**

* Oozie automation is implemented for extracting data from RDBMS into HDFS, as well as Hive warehouse table creation and partitioning (ETL).
* The process is coordinated to run every 20 minutes between 9-6pm Mon-Fri.
* **/ETL/ETL\_req\_2-2-3\_Oozie\_automation/**

**Oozie (Sqoop and Hive optimized)**

* Oozie automation is implemented for extracting new incremental data from RDBMS into HDFS, as well as loading that data into Hive warehouse.
* The process is coordinated to run every 20 minutes between 9-6pm Mon-Fri.
* **/ETL/** **ETL\_req\_2-2-4\_Oozie\_Incremental/**

**Visualization**

* Data is analyzed and visualized through the use of Hive queries and the Hive visualization tools,
* **/ETL/ Visualizations\_Explorations/**

**Pig(optional)**

* N/A