**2.1 [Customer and Transaction Details Module]**

Creating the database with MySQL.

PATH: /case\_study/db.zip/

1. Unzip db.zip found under the path above.
2. Ensure MySQL Server is running.
3. If not exists, create the MySQL connection with MySQL workbench with the following properties:
   * User: root
   * Password: password
   * Port: 3306 (Local Instance)

This connection is used throughout the requirements.

PATH: /*unzip\_path*/db/dumps/cdw\_sapp.sql

1. Open the existing .sql file named cdw\_sapp.sql with MySQL Workbench found under the path above.
2. Execute the file to automatically create and populate the database cdw\_sapp with records used during requirements.

Executing with Eclipse.

PATH: /case\_study/cdw\_workspace.zip/

1. Unzip cdw\_workspace found under the path above.

PATH: /*unzip\_path*/cdw\_workspace/

1. Set the Eclipse’s default workspace to the path above. This is where the project contents is located.
2. On first opening the project, you may experience an error, due to a missing JDBC jar. Follow these steps to add the jar:
   * Right on Referenced Libraries in the Eclipse file explorer and navigate to Build Path > Configure Build Path.
   * Navigate to Add External Jar.

PATH: /*unzip\_path*/cdw\_workspace/CDW\_SAPP/lib/mysql-connector-java-8.0.11

* + Use the provided file explorer to add the JDBC jar with the above path.

1. Execute the code by navigating to the green play button.

NOTES:

* Test data can be found in the form of comments in classes CustomerRunner and TransactionRunner respectively, under the package com.cdw.runner in the Eclipse file explorer.
* The exported .csv file can be found under the path /*unzip\_path*/cdw\_workspace/CDW\_SAPP/*output\_date\_time.*
* The .csv file is named with the full date and time (including seconds) at the time of export.

**2.2.1 [Data Extraction and Transportation with Sqoop]**

1. Open Hadoop’s local terminal on any web browser with the URL: *your\_vm\_ip*:4200.

PATH: /case\_study/Automation/Extract All Jobs/Extract All Jobs

1. Copy and paste each sqoop job command found in the path above into terminal and execute to create the sqoop jobs.

PATH: /case\_study/Automation/Extract All Jobs/Run

1. Copy and paste each sqoop run command found in the path above into terminal and execute to run the sqoop jobs.

NOTES:

* Path /user/maria\_dev/Credit\_card\_System/ in HDFS contains the extracted data.
* Notice that there are two different directories in path /case\_study/Sqoop/ named /Extract All Jobs/ and /Extract Inc Jobs/.
  + Extract All Jobs contains sqoop commands used for requirements 2.2.1 (Sqoop) and 2.2.3 (Oozie: Automation).
  + Extract Inc Jobs are used for requirement 2.2.4 (Oozie, Optimization). These sqoop jobs are incremental.
* The mapping logic can be found in the path /case\_study/Sqoop/mapping\_logic. This contains the queries used to transform the data from MySQL. This file is compatible with MySQL.

**2.2.2 [Data Loading Module]**

1. Open Ambari by opening any web browser with the URL: *your\_vm\_ip*:8888.
2. Enter login credentials for user maria\_dev and navigate to Hive View.

PATH: /case\_sudy/Hive/load\_hive

1. Copy and paste the contents of the file in the above path into hive and execute.

NOTES:

* Path /user/maria\_dev/Credit\_Card\_System/Part/ in HDFS contains records after partition manipulation.

**Prerequisite for Requirements 2.2.3 and 2.2.4**

Adding the java-json.jar.

PATH: /case\_study/java-json.jar

1. Place the above jar on HDFS path /user/oozie/share/lib/lib\_20161025075203/sqoop/

NOTES:

* The lib\_20161025075203 directory may be different, if this is the case, the workflows for requirements 2.2.2 and 2.2.3 need to be changed on the <archive> tag.

**2.2.3 [Automation Process with Oozie]**

PATH: /case\_study/Automation/job.properties.

1. Place the file located in the above path into package /root/Desktop/oozie/2.2.3/ on the linux virtual machine.

PATH: /case\_study/Automation/coordinator

PATH: /case\_study/Automation/workflow

1. Place the two files located in above paths into package /user/maria\_dev/oozie/2.2.3/ on HDFS.

PATH: /case\_study/Hive/load\_hive

1. Place the above file in the package /user/maria\_dev/load\_hive on HDFS.
2. Ensure step 2 of 2.2.1 [Data Extraction and Transportation with Sqoop] is complete.

PATH: /case\_study/Automation/Run

1. Execute the workflow with command found in the above file.

NOTES:

* The target folder for extracted records (/user/maria\_dev/Credit\_Card\_System/) is deleted in the “start-metasore” node.
* Database does not need to be dropped due to the “insert overwrite” clause.
* The <global> tag available in version 0.4 and up allows the nameNode, jobTracker, etcetera to be declared only once.

**Prerequisite for Requirement 2.2.4**

Transitioning to incremental imports.

PATH: /user/maria\_dev/Credit\_Card\_System/

1. Ensure the above folder in HDFS is deleted to prevent duplicate records.
2. Ensure database CDW\_SAPP in Hive is dropped.
3. Kill Oozie coordinator job created in 2.2.3 using the -kill argument.

**2.2.4 [Process Optimization Module]**

PATH: /case\_study/Optimization/job.properties.

1. Place the file located in the above path into package /root/Desktop/oozie/2.2.4/ on the linux virtual machine.

PATH: /case\_study/Optimization/coordinator

PATH: /case\_study/Optimization/workflow

1. Place the two files located in above paths into package /user/maria\_dev/oozie/2.2.4/ on HDFS.
2. Ensure step 3 of 2.2.3 [Automation Process with Oozie] is complete.

PATH: /case\_study/Sqoop/Extract Inc Jobs/create\_inc\_jobs.sh

1. Place the above script in the /root/Desktop/oozie/ on the linux virtual machine.

PATH: /case\_study/Sqoop/Extract Inc Jobs/run\_shell

1. Execute the above commands found in the file above to automate the sqoop job creations (create\_inc\_jobs).

PATH: /case\_study/Optimization/Run

1. Execute the workflow with command found in the above file.

NOTES:

* The above create\_inc\_jobs script had to be written in a single continuous line, else errors occurred. The file /case\_study/Sqoop/Extract Inc Jobs/Extract Inc Jobs contains the code in a human readable format.
* The script /case\_study/Sqoop/Extract Inc Jobs/delete\_inc\_jobs.sh is included for deleting sqoop’s incremental jobs and resetting the last values.
* *Refer to the last two notes in section 2.2.3 [Automation Process with Oozie].*

**2.2.5 [Data Visualization]**

PATH: /case\_study/Visualization/

***Additional Directories***

PATH: /case\_study/Requirements/

* contains the full details for the scope of the project.

PATH: /case\_study/Credit\_Card\_System.zip/

* contains an all files created during the Sqoop and Hive process. Expect similar result when executing.