**2.1 Core Java**

Core\_Java/

* JavaProject.zip -
  + 1. **RDBMS/mySQL data transformation**

RDBMS/

* *updated\_cdw\_data.zip* – updated database
* *sql.txt* - SQL server database, updated with LAST\_UPDATED TIMESTAMP column added to cdw\_sapp\_customer and cdw\_sapp\_brunch tables, so it can have updated timestamps every time data is modified or added to these tables.

**2.2.1 Data Extraction and Transportation with Sqoop**

Hadoop/HDFS/

* *Hadoop.txt* – Sqoop statement for Importing data from SQL server into HDFS

**2.2.2 Data Loading with Hive**

Hive/

* *Hive.txt* - script with Hive statements to create and populate (with data extracted in previous step) hive tables in Hadoop Filesystem.

**2.2.3 Automating Sqoop and Hive processes from steps 2.2.1 and 2.2.2 with Oozie Workflow.**

Oozie\_1/

* *workflow.xml* – oozie xml script for extracting and modifying data based on requirements with Sqoop, creating and populating tables with Hive using scripts.
* *external\_creditcard.txt* – script with hive statements to drop, create and populate cdw\_sapp\_creditcard with extracted data.
* *external\_branch.txt* - script with hive statements to drop, create and populate cdw\_sapp\_branch with extracted data.
* *external\_customer.txt* - script with hive statements to drop, create and populate cdw\_sapp\_customer with extracted data.
* *external\_time.txt* - script with hive statements to drop, create cdw\_sapp\_time.
* *insert\_time.txt* - script with hive statements to populate cdw\_sapp\_time with modified data from cdw\_sapp\_creditcard table.
* *job.properties* - configuration file, containing all parameters needed to run Oozie workflow.
* *coordinator.xml* – Oozie workflow scheduler.

**2.2.4 Process Optimization Module Oozie (Sqoop and Hive optimized)**

Oozie\_2/

* *workflow\_2.xml* – updated Oozie workflow script for extracting only *new data* using Sqoop Job and updating original tables with new data using Hive.
* *sqoop\_jobs.txt* – script with Sqoop statements that needs to be run from terminal before running Oozie Workflow.
* *create\_base\_tables.txt* - script with hive statements to drop if exists and create cdw\_sapp\_branch, cdw\_sapp\_customer, cdw\_sapp\_creditcard, cdw\_sapp\_time. cdw\_sapp\_creditcard, cdw\_sapp\_time tables will also be populated during this step, as this tables/data will remain as it is (no unique keys available for creditcard tables, not possible to update). This script will only run once, when oozie job executes first time. Then this step will be skipped based on the <decision> action.
* *insert\_into\_incremental\_tables.txt* - script with hive statements to create and populate incremental tables for cdw\_sapp\_branch, cdw\_sapp\_customer with new data.
* *merge\_with\_new\_data.txt*- script with hive statements to update originals cdw\_sapp\_branch and cdw\_sapp\_customer tables with new data.
* *job.properties* - configuration file, containing all parameters needed to run Oozie workflow.
* *coordinator.xml* – Oozie workflow scheduler.

**2.2.5 Data Visualization**

*Hive\_visualization.docx* - hive queries and screen shots of the graphs.