1. **Mini Project 1 - Prediction Airline Delay with Hive and Logistic Regression**

* ***Technical Framework and Skills:***

Linux Shell Script, SQL, HDFS, Yarn, Tez, Hive and Hive machine learning framework (Hivemall)

* ***Major Assignments:***

1. Ingested the data from file system into HDFS.
2. Create external hive tables over the dynamic partition datasets in HDFS.
3. Filtered, projected, transformed and joined different hive tables.
4. Extracted 11 features from the dataset.
5. Processed, Derived, transformed with hive scripts and custom UDF functions.
6. Normalize the features with min-max, zscore and standard deviation.
7. Built and trained the model with Logistic Regression and hivemall UDF, UDAF functions.
8. Computed the performance metrics and evaluated the performance of Logistic Regression model.
9. Optimized hive join and tuned the hive query performance.

1. **Mini Project 2 - Food Mart Sales Data Analytics with Apache Kylin OLAP and CUBE, Hive, HBase**

* ***Technical Framework and Skills:***

SQL, HDFS, Yarn, MapReduce, Apache Kylin, Hive, HBase, Oozie, BI tools as Excel or Tableau over Kylin ODBC, REST API

* ***Major Assignments:***

1. Set up Apache Kylin cluster (Query Engine, CUBE Build and Design Server) and integrated them with Hadoop Cluster (Hive, HBase cluster)
2. ETL the raw datasets into Hive tables
3. Created sales fact table and customer, product, inventory, store lookup tables from Hive tables.
4. Designed dimensions, measure (MAX, TOP N, Count) per brand name and sales region city, sales model and cube.
5. Optimized the sales cube build process using derived column, hierarchy, aggregation groups and partitioned column, row key design.
6. Versioned the sale cube metadata and packaged and deployed the sales cube for different environments.
7. Developed and launched Oozie workflow jobs to automate the sales cube build process and persist the pre-aggregation data into HBase.
8. Integrated cube query with BI tools and Linux shell script by Kylin ODBC driver and REST API call.
9. **Mini Project 3 - End to End Streaming stocks price CUBE build with Apache Kylin OLAP and streaming CUBE, Kafka, Apache Nifi and HBase**

* ***Technical Framework and Skills:***

SQL, Kafka, Nifi, Yarn, MapReduce, Apache Kylin and REST API

* ***Major Assignments:***

1. Configured Kafka cluster and created partitioned topic
2. Developed the apache nifi workflow to fetch the stocks price from google finance, extracted, transformed and split raw JSON messages, and publish the enriched JSON message to Kafka topic, which is being listened by Kylin CUBE adaptor.
3. Mapped Kafka JSON message with Apache Kylin data model, establish the timestamp parse strategy and choose partitioned column for the windowed streaming data ingestion into CUBE.
4. Designed dimensions, measure (max and total price for each stock during the window of streaming stock price data) per exchange and ticker, stock price model and cube.
5. Worked on the solution to use apache Nifi workflow to automate the streaming stocks price cube build process and persist the pre-aggregation data into HBase.

1. **Mini Project 4 - Prediction Airline Delay with Datameer and Decision Tree**

* ***Technical Framework and Skills:***

HDFS, Datameer, Yarn, Tez and Datameer machine learning, Avro

* ***Major Assignments:***

1. Installed Datameer at one of edge nodes in the hadoop cluster.
2. Explored the different datameer jobs execution engines such as spark and tez.
3. Created the connection to retrieve the flight and weather datasets from HDFS.
4. Scheduled the import jobs to ingest the flight and weather data from HDFS.
5. Worked on the workbooks design to automate the process of datasets generated from import jobs via a chain of workbook sheets transformation, filtering, sorting the workbook sheet, performing the aggregation by groups, joining multiple workbook sheets.
6. Developed custom datameer function using plugins to support the features extraction.
7. Learned to use datameer decision tree to train the workbook sheet with selected features and predict the result from the test dataset.
8. Created the export jobs to export the predicted flight delay and performance metrics into HDFS as avro format once the workbooks completed the calculation and prediction.
9. Applied the generated datasets to a group of infographics widgets on the dashboard.
10. Integrated the datameer with the apache atlas for data governance via datameer event bus, internal kafka topic and atlas API.

1. **Mini Project 5 - ETL Registered Business Locations from RDBMS into HDFS using Sqoop, Oozie and Hive**

* ***Technical Framework and Skills:***

HDFS, MySQL, Yarn, MapReduce, Sqoop and Oozie, Parquet, Linux shell scripting

* ***Major Assignments:***

1. Created the Linux shell script to grab the business dataset from external website periodically.
2. Uploaded the business dataset into MySQL using MySQL script.
3. Implemented the sqoop job to incrementally import the structure data into HDFS which is organized and partitioned by the current date and stored as parquet format, and last value is automatically saved into sqoop meta data store
4. Create external hive table to refresh the newly created partitioned dataset into Hive meta data store
5. Practiced the Sqoop best approach on how to evenly split the datasets into mappers in parallel, reduce the data skew chances and protected the password for RDBMS access using password alias and hadoop credentials loader, data type mapping between RDBMS and hive.
6. Weaved the actions or scripts above into the ETL pipeline using oozie workflow, scheduled and executed them using oozie coordinator.