# Internship Assessment I

**Name-Ankit Yadav**
**Employee ID-10201**
**Mentors-Chetan Khatri**
**& Babu Prabhakar**

## Assigned Tasks:

1. Install Hortonworks Distribution 2.6.5 on VirtualBox/VMware
   i. Ambari UI
   ii. Spark UI(along with Spark History Server)
   iii. YARN UI

2. Install Airflow and work with following Operators:
   i. Bash Operator
   ii. Python Operator
   iii. SparkSubmit Operator

3. Spark with HDFS
   i. Use partitionBy for Categorical data(Parquet)
   ii. Use HDFS to read and write
   iii. Write hdfs data into db tables

4. Difference between Parquet and AVRO.

5. Work with .yaml and .config files.

## Hortonworks:

- HDP includes Apache Hadoop and is used for storing, processing, and analyzing large volumes of data.

- The platform includes Hadoop technology such as the Ambari, Hadoop Distributed File System, MapReduce, Spark, Pig, Hive, HBase, ZooKeeper, Oozie and additional components.

  - Ambari UI(8080):

    An open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters.

  - Spark UI(4040):

    Web UI is the web interface of a Spark application to monitor and inspect Spark job executions in a web browser.
    (SparkContext object is destroyed then can’t access Spark UI)

  - Spark History Server(18080):

    Spark History Server is the web UI for completed and running Spark applications. It is an extension of Spark’s web UI.
    (You can see the UI also after SparkContext variable is destroyed)

  - YARN UI(8088):

    Is a resource manager UI ,similar to hadoop UI.

**Tasks Performed:**

- Ran spark jobs in Hortonworks shell.
- Went through the Ambari UI

**Problem Statement:** While getting started with HDP, I faced problem in installation phase. System Hanged up every time on installation.

**Solution:** By Extending RAM to 16GB ot worked fine.

# Airflow:

- Airflow is a platform to programmatically author, schedule and monitor workflows.

- Can schedule cron jobs.

- Airflow workflow is designed as a directed acyclic graph (DAG).The airflow scheduler executes your tasks on an array of workers while following the specified dependencies.

- The user interface makes it easy to visualize pipelines, monitor progress, and troubleshoot problems when needed

BashOperator: Executes commands as bash_script.(Example Code)

PythonOperator: Executes Python Callables.(Exaple_Code)

SparkSubmitOperator: Executes Spark jobs.(Example_Code)

**Tasks Performed:**

- Scheduled spark jobs using BashOperator and SparkSubmitOperator.

- Ran complex DAGs and observed the Airflow UI

**Problem Statement:** While working with Airflow i was stuck at SparkSubmit Operator and other one was whenever i tried to run larger spark jobs than airflow server gets down.

**Solution:** By adding a plugin file(spark_operator_plugin.py) from DataBricks SparkSubmitOperator worked fine.

## Working with AVRO and Parquet:

**Avro:**

- Avro is a row-based storage format for Hadoop.
- Avro stores the data definition in JSON format making it easy to read and interpret.
- AVRO can change schema over time, e.g. adding or removing columns from a record.
- Avro stores both the data definition and the data together in one message or file.
- If you want to scan or retrieve all of the fields in a row in each query, Avro is usually the best choice.

**Parquet:**

- Parquet is a column-based storage format for Hadoop.
- If your dataframe has many columns, and you want to work with a subset of those columns rather than entire records, Parquet is optimized for that kind of work.

**Problem Statement:** No built-in package for AVRO file format in spark.

**Solution:** By adding package from DataBricks it ran successfully.

(Same thing worked for xml file also)

**Tasks Performed:**

- Read and Write files in HDFS in AVRO and Parquet format.(Example Code)

- Use partitionBy to categorize the parquet data.(Example Code)

## Yaml File:

- Human Readable File.

- Easy to code and provides powerful configuration settings.

**Tasks Performed:**

- Performed database operation using yaml file.
  (Example code for .yaml)
  (Example Code for use of yaml)