**Lab : Batch Updates to a S3 Datalake using Apache Hudi**

**Copy on Write**

**Overview**

This lab demonstrates using PySpark on [Apache Hudi](https://aws.amazon.com/emr/features/hudi/) on Amazon EMR to insert/upsert/delete records to an S3 data lake.

This lab covers the following concepts when writing Copy On Write tables to an S3 Datalake:

* Write Hudi Spark jobs in PySpark.
* Bulk Insert the Initial Dataset.
* Write a Non-Partitioned table.
* Sync the Hudi tables to the Hive/Glue Catalog.
* Upsert some records to a Hudi table.
* Delete records from a Hudi table.
* Rollback.

**Please follow the steps listed below to complete this LAB**

**Step-1: Create EC2 Key Pair**

1. Navigate to the EC2 console
2. Navigate to Key Pairs under ‘Network and Security’
3. Click on ‘Create Key Pair’. Provide an appropriate name for the Key Pair
4. Download the generated .pem file and keep it at a secure location. If you use a Windows machine, you can use puttygen to convert the .pem file to a .ppk file.

**Step-2 : EMR Cluster Creation**

**This Lab is based on Hudi version 0.5.0-incubating and runs fine on Jupyter Notebooks connected to a 1 node (r5.4xlarge) EMR cluster with configuration listed below**

* EMR versions 5.29.0 or 6.0.0
* Software Versions
* - Hadoop 2.8.5
* - Hive 2.3.6
* - Livy 0.6.0
* - JupyterHub 1.0.0

- Spark 2.4.4

* AWS Glue Data Catalog settings - Select the below listed check boxes

- Use for Hive table metadata

- Use for Spark table metadata

**Refer to the below link for detailed instructions to create an EMR cluster**

**https://emr-etl.workshop.aws/en/cluster\_creation/cluster.html**

**Step-3 :Connect to the Master Node of EMR cluster Using SSH :**

- ssh -i ~/xxxx.pem hadoop@<ec2-xx-xxx-xx-xx.us-west-2.compute.amazonaws.com>

**- Ensure the below listed files are copied into HDFS.**

- hadoop fs -copyFromLocal /usr/lib/hudi/hudi-spark-bundle.jar hdfs:///user/hadoop/

- hadoop fs -copyFromLocal /usr/lib/spark/external/lib/spark-avro.jar hdfs:///user/hadoop/

- hadoop fs -copyFromLocal /usr/lib/spark/jars/httpclient-4.5.9.jar hdfs:///user/hadoop/

**Step- 3: Create EMR-Notebook**

* Follow the instructions at the link provided below to create an EMR NoteBook

https://emr-etl.workshop.aws/en/emr\_notebooks\_sagemaker/create\_emr\_notebook.html

**Step- 4 : Execute the LAB**

* Upload the LAB-COW.ipynb file to Notebook and follow along the instructions in the notebook to complete the lab.