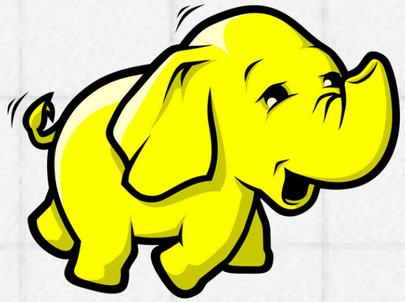


Capstone Project -3

Presented by Soumya S. Panda

Dynamic Data Ingestion and Storage in HDFS with Automated Hive Integration

This project demonstrates the process of ingesting a CSV file into Hadoop's HDFS and then using Apache Hive to create a database and table, load the CSV data, and query it.

Project Flow

01. Upload CSV Data to HDFS

02. Create Hive Database, Load Data, and Query Results

03. Automate the process

Project Execution

The project is organized into two main scripts :

Hadoop Bash Script

HiveQL Bash Script

Hadoop Bash Script

save_to_hdfs.sh --> Script to upload CSV data to HDFS

Bash Script

=> chmod +x save_to_hdfs.sh

=>./save_to_hdfs.sh

HiveQL Bash Script

hive_operations.sh -->Script to create Hive database and table, load data, and query it

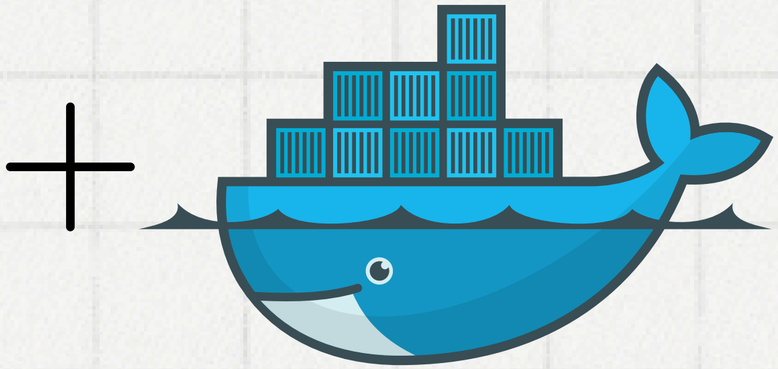
Bash Script

=> chmod +x hive_operations.sh

=>./hive_operations.sh

Project Implementation

DevBox



Data Fusion

Docker

AWS EC2

Prerequisite

. **Installing DevBox with Docker Into EC2 Instance VM**

. **Hadoop Services (HDFS & YARN) Must be Running**

. **Data files should be availabe into the local code env**

Thank you very much!