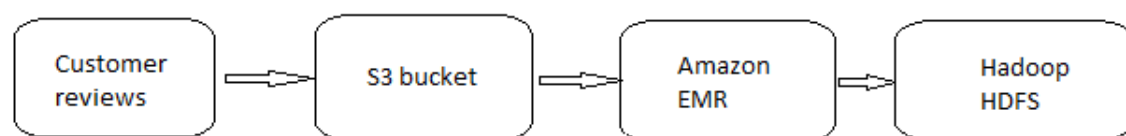# Sentiment Analysis

---

Wireframe Documentation

We have divided our project into two sections:
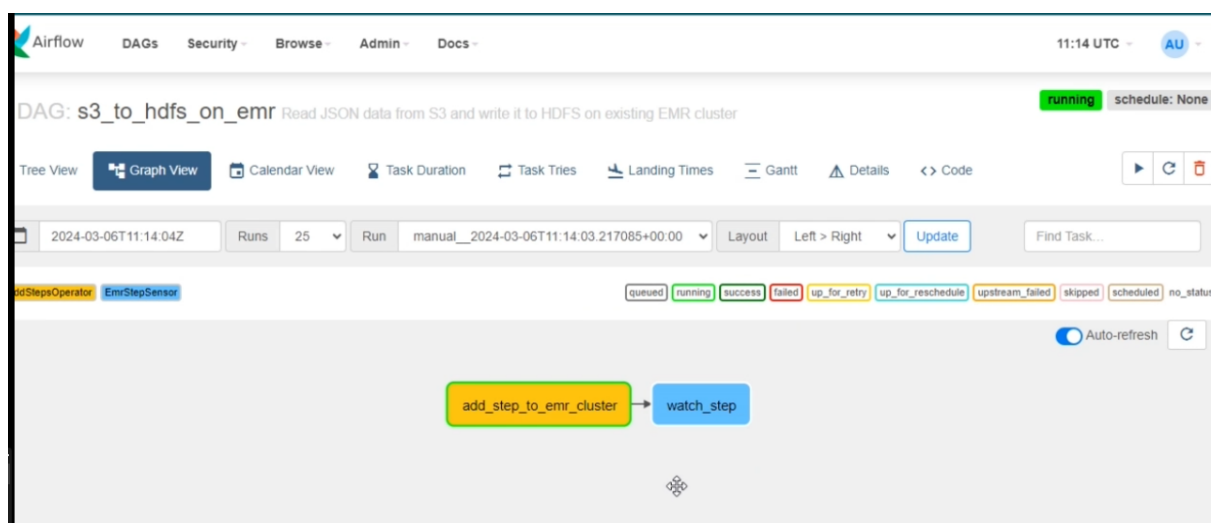
First section include-

Q1) How to make scalable pipeline using spark to read customer reviews and store it into HDFS

We have customer reviews which is stored in S3 bucket in Json format. We created a EMR cluster which create the ec2 instance for Hadoop cluster in which we store the review in HDFS in parquet file format

Q2) How to schedule the pipeline to run iteratively after each hour.

To iterate the pipeline, we can use Aws eventbridge or external trigger such as Airflow. In this project, I have created a Dag to trigger the pipeline

In second section –

Q1) How to do sentiment analysis using spark machine learning

```
+------------------------------------------+
|                                          |
|             HDFS                         |
| (Customer Review Data Repository)        |
|                                          |
+------------------------------------------+
                  |
                  |
                  v
+------------------------------------------+
|                                          |
|           Apache Spark                   |
|   (Distributed Data Processing Engine)   |
|                                          |
+------------------------------------------+
                  |
                  v
+------------------------------------------+
|                                          |
|        Spark Machine Learning            |
|          (Logistic Regression)           |
|                                          |
+------------------------------------------+
                  |
                  v
+------------------------------------------+
|                                          |
|        Sentiment Analysis Results        |
|                                          |
+------------------------------------------+
```