High Level Design Document (HLD): Sentiment Analysis Pipeline Using HDFS and Spark

CONTENTS:

* Abstract………………………………………………………………..3
* Introduction………………………………………………………….4
* Problem Statement……………………………………………….4
* Architecture…………………………………………………………..5
* Design Details………………………………………………………..5
* Benefits………………………………………………………………….6
* Conclusion……………………………………………………………..7

**Abstract**

Sentiment analysis and textual emotion recognition are closely related. Sentiment analysis is target-oriented, aiming to identify opinions or attitudes towards topics or entities (e.g., product, movie, Customer Reviews etc…). Emotion recognition, on the other hand, focuses on recognizing either the emotion expressed in text or evoked by the text, with no attachment to a specific target.

Sentiment analysis or opinion mining is the computational study of people’s opinions, sentiments, attitudes and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years.

**Introduction**

Our goal is to develop a scalable and efficient pipeline for performing sentiment analysis on customer reviews using Spark. The pipeline will be designed to read data from an S3 bucket, and store it in HDFS, and then need to perform sentiment analysis on the data using Spark Machine Learning. The output of the above analysis will be stored back in HDFS for further analysis.

**Problem statement**

Design scalable pipeline using spark to read customer review from s3 bucket and store it into HDFS. Schedule your pipeline to run iteratively after each hour.

Create a folder in the s3 bucket where customer reviews in json format can be uploaded. The Scheduled big data pipeline will be triggered manually or automatically to read data from The S3 bucket and dump it into HDFS.

Use Spark Machine learning to perform sentiment analysis using customer review stores in HDFS.

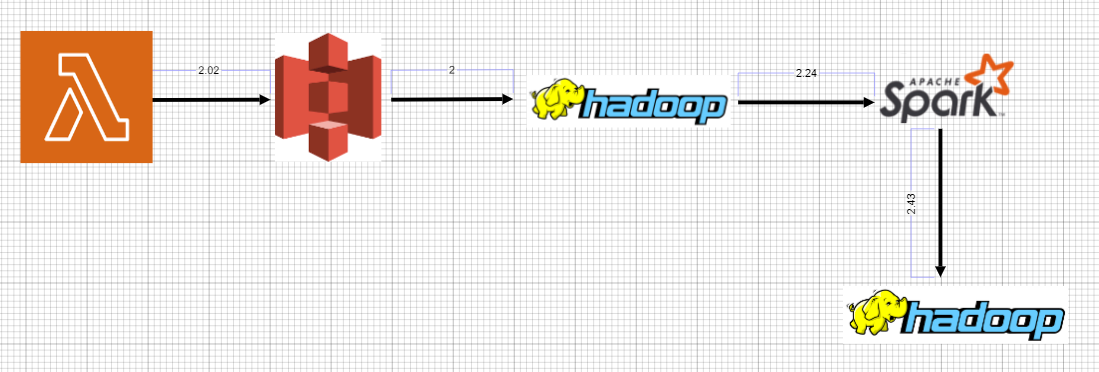
**Architecture**

The architecture of the proposed pipeline consists of three main components:

Data Source: The pipeline will read customer reviews data from an S3 bucket. This data will be in JSON format.

Processing Engine: Spark will be used to process the data, perform sentiment analysis using ML libraries, and store the output in HDFS.

Storage: The final output of the pipeline, which includes sentiment analysis results, will be stored in HDFS.



**Design Details**

Data Source: The pipeline will read customer reviews data from an S3 bucket. This data will be in JSON format.

We will upload a file to s3 which is in csv format. By using Lambda functions we are changing the format from csv to json and saving the json file in different folder in same s3 bucket.

For every 1 hour lambda function will get triggered by event bridge.

**Processing Engine:**

Once the data is uploaded to the S3 bucket, the pipeline will be triggered manually or automatically to read data from the S3 bucket and dump it into HDFS. We will use Spark to perform this operation.

The data will be processed using Spark, which is a fast and distributed processing engine for big data. We will use Spark’s Machine Learning Library (ML libraries) for performing sentiment analysis on the customer reviews.

To execute the pipeline, we will create a Spark job in Python. The job will be responsible for reading data from the S3 bucket, performing sentiment analysis using ML libraries, and storing the output in HDFS.

**Storage:**

The final output of the pipeline, which includes sentiment analysis results, will be stored in HDFS. The output can be analysed further using other big data tools such as Apache Hive, Hue…

**Benefits**

**Accuracy**: The use of Spark’s Machine Learning Library ensures accurate results.

**Real-time analysis**: The pipeline can be scheduled to run iteratively after each hour, which ensures real-time feedback based on the customer satisfaction.

**Scalability**: The proposed pipeline is highly scalable and can handle large volumes of data.

**Efficiency**: The use of Spark enables the pipeline to process and analyse the data efficiently, which leads to faster results.

**Reusability**: The code written in the components used should have the ability to be reused with no problems

**Application Compatibility**: The different components of the project will be using python as interface between them. Each component will have its own task to perform, and is the job of the python to ensure proper transfer of information.

**Resource Utilization:** when any task is performed, it will likely used all the processing power available until that function is finished.

**Conclusion**

Finally conclusion, The pipeline uses Spark to process the data, perform sentiment analysis using ML libraries, and store the output in HDFS. This pipeline provides real-time feedback on customer satisfaction and enables businesses to identify for improvements and analysis on the future. It also help to identify the imbalance and can take the necessary action on the errors/corrections to stop or rectify based on the customers. Spark-based Sentiment Analysis Pipeline for Customer Reviews is a scalable and efficient solution for analysing customer reviews data.