Low Level Design Document (LLD): Sentiment Analysis Pipeline Using HDFS and Spark

CONTENTS:

* Introduction………………………………………………………….3
* Architecture………………………………………………………….3
* Explanation……………………………………………3
* Tools used…..………………………………………………………..4
* Conclusion…………………………………………………………….4

**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.

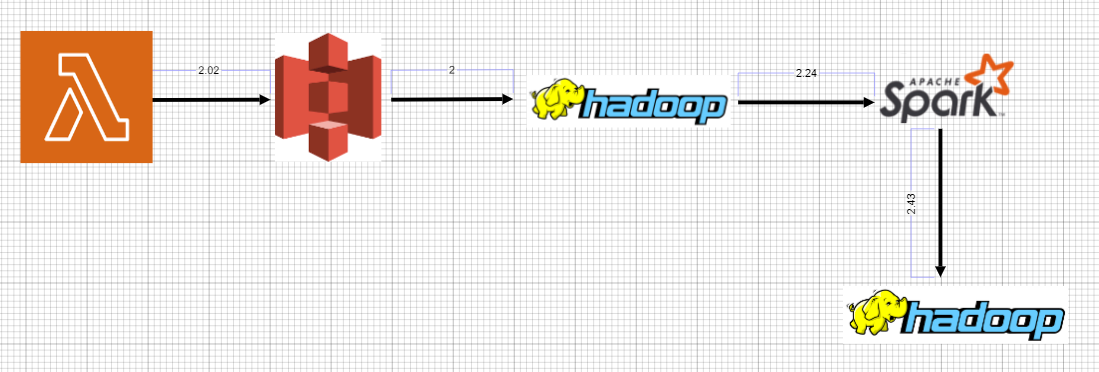
**Architecture**

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.



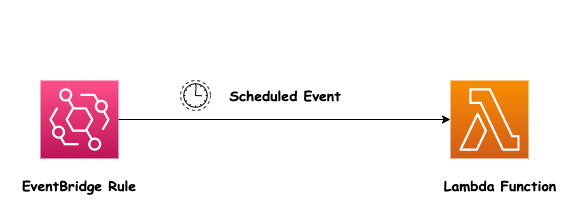
Explanation:

AWS Lambda will read data/file from S3 by Event Notification configured on the s3 bucket. The Lambda function will read the CSV file(sentiment-anal-intern26/data.csv, convert the data it to JSON format, and store the resulting file in the sentiment-anal-intern26/jsonfile/data.json.

Added Event bridge to Lambda function, it will triggered every 1 hour. If any new file is detected it will start the process again. From Hadoop the Apache Spark will read the input file to perform analysis.

Spark machine learning will be used to perform sentiment analysis on the customer reviews. The resulting data will be stored in a new folder in HDFS.

**Tools Used**



**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.