High Level Design Document: Spark-based Sentiment Analysis Pipeline for Customer Reviews

CONTENTS:

|  |  |
| --- | --- |
|  | Page Number |
| **Introduction** | 1 |
| **Architecture** | 1 |
| **Detailed Procedure** | 2 |
| **Benefits** | 2 |
| **Conclusion** | 3 |

1. **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, store it in HDFS, and perform sentiment analysis using Spark Machine Learning. The output of the pipeline will be stored back in HDFS for further analysis.

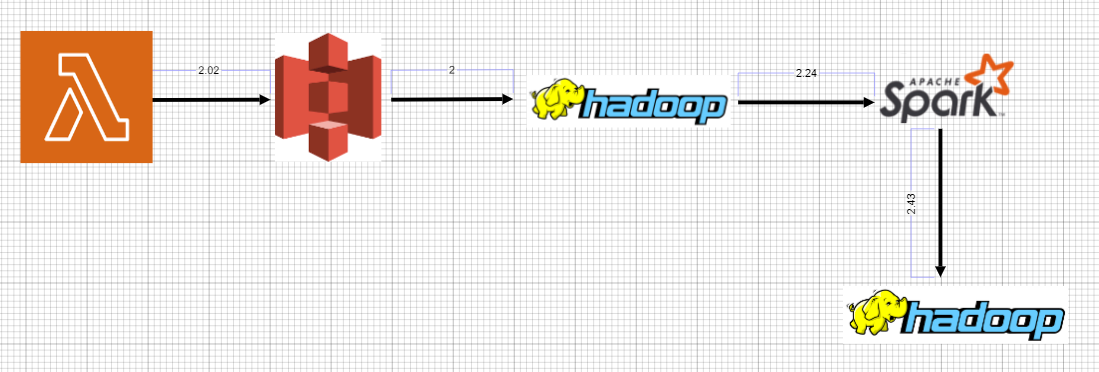
1. **Architecture:**

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 MLlib, and store the output in HDFS.

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

Image:  


1. **Detailed Explanation:**

**Data Source:**

The data source for our pipeline is an S3 bucket where customer reviews in CSV format can be uploaded.

A separate folder will be created in the S3 bucket for storing customer reviews data in JSON format.

**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 (MLlib) for performing sentiment analysis on the customer reviews.

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

This project has been done in manual mode. To automate this pipeline, we have explained the above point

**Storage:**

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

1. **Benefits:**

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

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

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

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

1. **Conclusion:**

In conclusion, the proposed Spark-based Sentiment Analysis Pipeline for Customer Reviews is a scalable and efficient solution for analyzing customer reviews data. The pipeline uses Spark to process the data, perform sentiment analysis using MLlib, and store the output in HDFS. This pipeline provides real-time feedback on customer satisfaction and enables businesses to identify areas for improvement.