Low Level Design

Sentiment Analysis

Contents

|  |  |
| --- | --- |
|  | Page Number |
| **Introduction** | 1 |
| **Tech Stack Used** | 1 |
| **Architecture** | 1 |
| **Detailed Procedure** | 2 |
| **Conclusion** | 2 |

1. **Introduction:**

The aim of this document is to provide a low-level design for a scalable pipeline that uses Spark for sentiment analysis. The pipeline will read customer reviews in JSON format from an S3 bucket, store it in HDFS, and use Spark machine learning to perform sentiment analysis.

1. **Tech Stack Used:**

The following technologies will be used in the pipeline:

Apache Spark

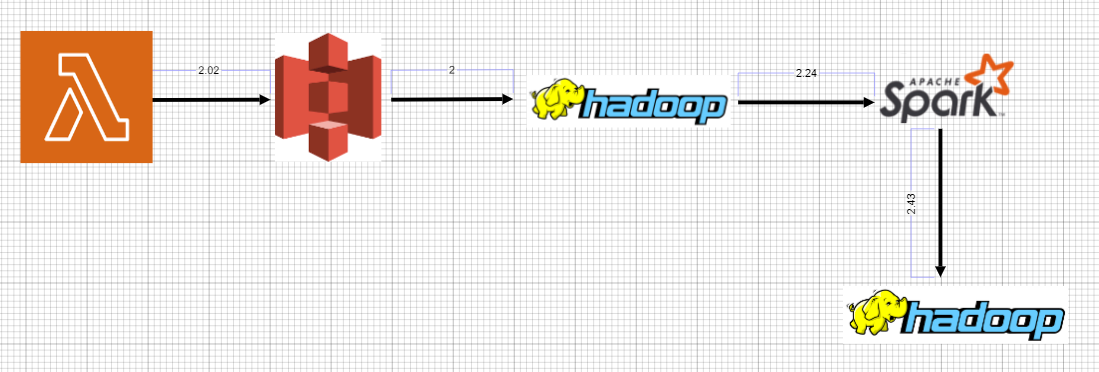
Amazon S3

Amazon Lambda

Hadoop Distributed File System (HDFS)

1. **Architecture:**

The architecture of the scalable pipeline is as follows:



**Architecture Explanation:**

AWS Lambda will be triggered by an S3 Event Notification configured on the internship/csv\_input/ S3 bucket. The Lambda function will read the CSV file, transform it to JSON format, and store the resulting file in the internship/json\_output/ bucket

Customer reviews will be uploaded in JSON format to an S3 bucket. A new folder will be created in the bucket to upload the reviews.

The pipeline will be triggered manually or automatically to read data from the S3 bucket using Apache Spark.

Apache Spark will use Hadoop to store the customer reviews in HDFS.

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.

1. **Detailed Procedure:**

The following steps will be taken to implement the scalable pipeline:

* Create an S3 Event notification triggered by an S3 bucket event.
* Lambda function will get triggered and convert the csv into json
* Create a new folder in the S3 bucket to upload customer reviews in JSON format.
* Create an Amazon EMR cluster with Apache Spark and Hadoop installed.
* Create a script that can be triggered manually or automatically to run the pipeline.
* The script will use the Apache Spark API to read customer reviews from the S3 bucket in JSON format.
* The customer reviews will be stored in HDFS using Hadoop Distributed File System.
* 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.
* Schedule the Pipeline(to automate the pipleline):

Use the cron job to schedule the pipeline to run iteratively after each hour.

Configure the cron job to run the script automatically.

1. **Conclusion:**

This low-level documentation provides a step-by-step guide for designing a scalable pipeline using Spark to read customer reviews from an S3 bucket and store it into HDFS. The pipeline uses Spark machine learning to perform sentiment analysis on the customer reviews. By scheduling the pipeline to run iteratively after each hour, the analysis will be up-to-date, and real-time feedback on customer satisfaction can be provided.