LOW LEVEL DESIGN

Sentiment Analysis

CONTENT

1.INTRODUCTION

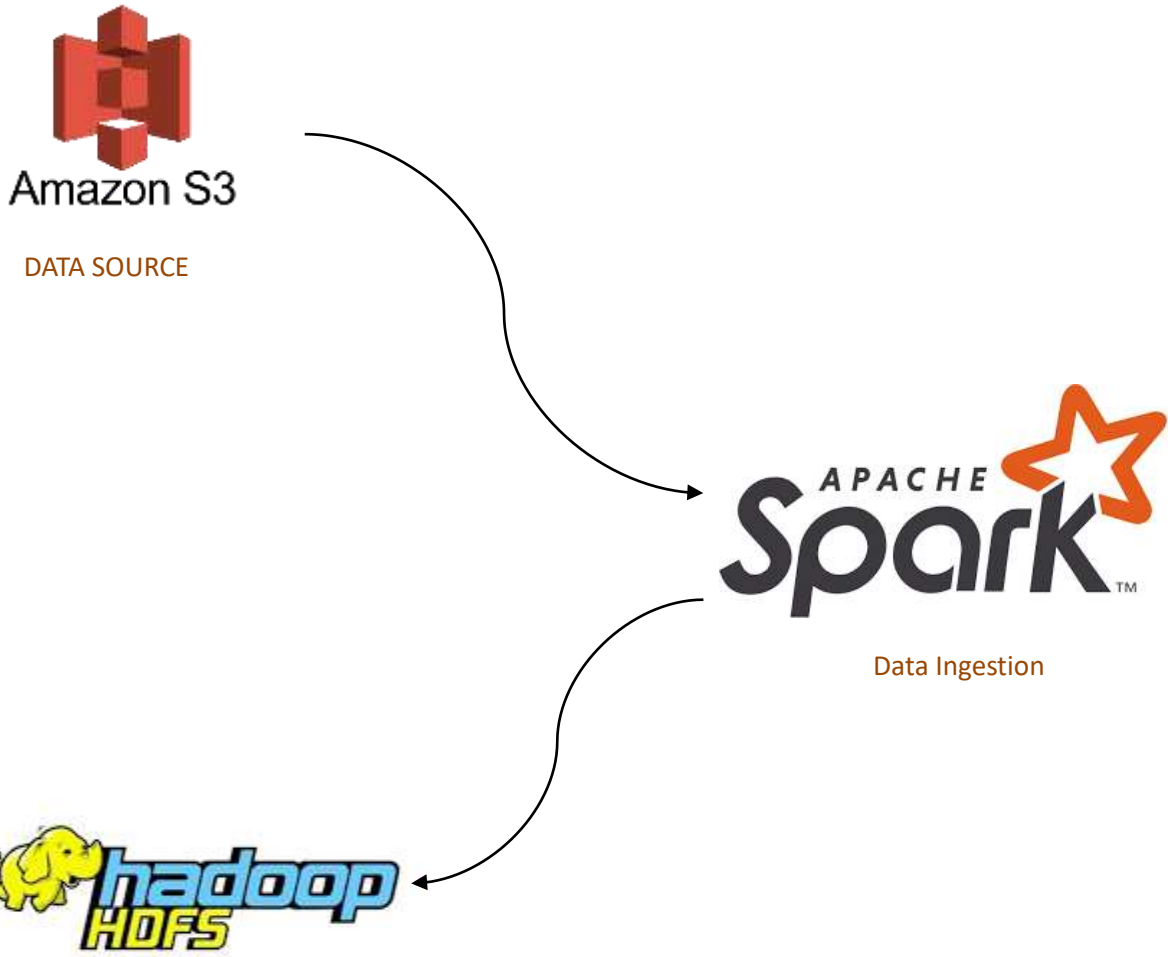
This project aims to develop a highly scalable and efficient data processing pipeline using Apache Spark, a distributed computing framework. The primary objective of this pipeline is to extract insights from a large volume of customer reviews by reading them from an S3 bucket and storing them into Hadoop Distributed File System (HDFS). By leveraging the power of Spark's parallel processing capabilities, the pipeline ensures timely and accurate analysis of customer feedback.

The availability of customer reviews in today's digital landscape is of paramount importance for businesses to make informed decisions and enhance customer satisfaction. However, the sheer volume of reviews poses challenges in terms of data management and analysis. To address this, the pipeline incorporates scheduled iterative execution, with a frequency of every hour, ensuring continuous data ingestion and processing.

To facilitate the seamless uploading of customer reviews, a dedicated folder is created within the S3 bucket. This folder serves as the source from which Spark retrieves the reviews for further processing. By leveraging Spark's distributed computing capabilities, the pipeline efficiently retrieves and transfers the customer reviews to HDFS, providing a reliable and scalable storage solution.

One of the key features of this pipeline is the utilization of Spark's machine learning capabilities for sentiment analysis. Sentiment analysis allows businesses to gain valuable insights into customer satisfaction levels, enabling them to identify areas for improvement and make data driven decisions. By employing Spark's parallel processing and in-memory computing capabilities, sentiment analysis can be performed in real-time or near-real-time, ensuring timely insights for actionable decision-making. The scalability of the pipeline is a critical consideration, given the exponential growth of customer review data. Spark's ability to distribute computation across multiple nodes and leverage in-memory caching enables efficient processing of large volumes of data. This ensures that the pipeline can handle increasing data loads without compromising performance

2.ARCHITECTURE



3.ARCHITECTURE DISCRIPTION

3.1 Data Source

- Customer reviews are stored in the Amazon S3 bucket location: "s3://sentimental-analysis-project-ineuron/customer_reviews.json".
- The customer reviews are in JSON format.

3.2 Apache Spark Cluster

- The pipeline utilizes an Apache Spark cluster with version 3.2.1 to perform distributed data processing and analysis.
- The Spark cluster comprises a master node and multiple worker nodes, enabling parallel execution of tasks.

3.3 Data Ingestion

- Spark reads the customer reviews data from the specified S3 bucket location, "s3://sentimental-analysis-project-ineuron/customer_reviews.json".
- The data ingestion process efficiently retrieves the customer reviews from S3 and loads them into Spark's distributed memory.

3.4 Storage in HDFS:

- The processed customer reviews, along with the sentiment analysis results, are stored in HDFS.
- The HDFS location for storing the processed data is "hdfs://localhost:9000/customer_reviews".
- Hadoop version 3.3.1 is used for managing the HDFS storage system.
- HDFS provides fault-tolerant and scalable storage, ensuring the persistence of the processed reviews for further analysis and access.