OPCITO TECHNOLOGIES

Optimizing Data Pipeline Using Big Data Analytics Techniques

About The Customer

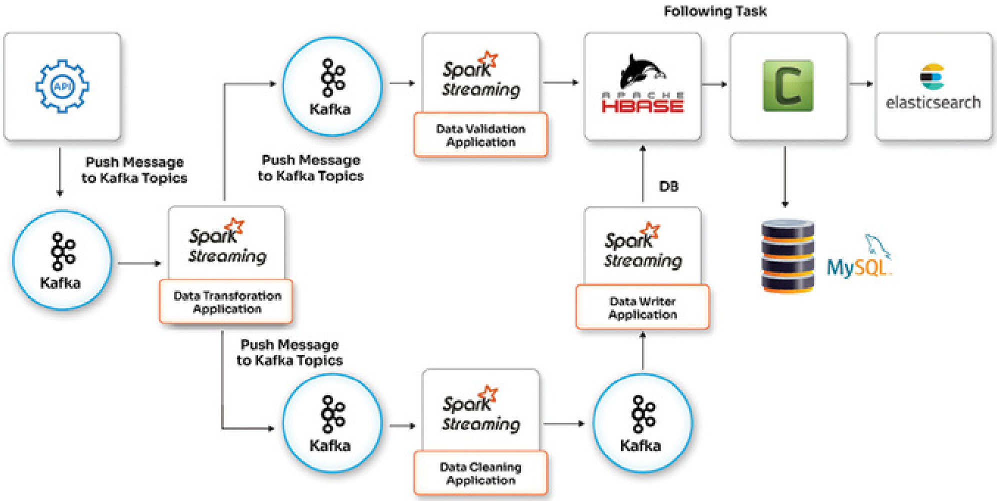
The client is a leading business, data analytics, and marketing automation service provider based in the US. These services enable businesses to understand various business operations, customer needs, future requirements, customer retention, and recommendation with predictive analysis to help them build online reputation and reviews.

Business Challenge

The client has a solution that provides their clientele with insights that can be used to build focused marketing campaigns based on the latest customer preferences. For this, the solution uses data from various data logs like CRMs and provides relevant insights with the help of data transformation and data processing activities. The problem with the previous mechanism used by the solution was the time consumed to process data. It was sequentially processing the received data using a traditional framework consisting of MySQL. This resulted in increased time to process data and produce results from a few hours to a few days, depending on the size of the records to be processed.

How Opcito Helped

After due brainstorming, the Opcito team decided to revamp the existing system by building a real-time data pipeline using Spark, Kafka, and HBase. Opcito's team of developers and architects designed a workflow that optimized resource utilization and minimized data processing steps and time. Our team of developers wrote a Spark streaming application with three jobs. The First Job (Transformation Job) was designed to manage the transformation of fetched records. The second job (Cleaner Job) managed the responsibility of data cleaning. The third job (Validator Job) was designed to filter out data based on validation rules. Once the records were processed and filtered through all three stages/jobs, the filtered data was written into HBase. Throughout the entire pipeline, Kafka was used for inter-job communication. Once the data was pushed into HBase, the celery worker woke up every few seconds and checked for the total number of records in HBase against the entries in MySQL. In case of a matched number of records, it will start dumping data to Elasticsearch for further analytics and into MySQL to allow for the traditional approach for the client's use.


Following Task
API
Push Message to Kafka Topics
Kafka
Push Message to Kafka Topics
Kafka
Spark Streaming
Data Validation Application
APACHE HBASE
C
elasticsearch
Spark Streaming
Data Transforation Application
DB
Spark Streaming
Data Writer Application
MySQL
Push Message to Kafka Topics
Kafka
Spark Streaming
Data Cleaning Application
Kafka

Technologies, Tools, and Platforms used

- SPARK
- KAFKA
- HBASE
- YARN 2

Benefits

SPEED	Faster delivery with the ability of parallel processing of multiple CRM data
PERFORMANCE	Performance optimization with the gilt-edge utilization of the computing power/resources available
COST SAVING	Reduced power consumption and computing power along with operating expenses

About Opcito

At Opcito, we believe in designing transformational solutions for our customers, start-ups, and enterprises, with our ability to unify quality, reliability, and cost-effectiveness at any scale. Our core work culture focuses on adding material value to your products by leveraging best practices in DevOps, like continuous integration, continuous delivery, and automation, coupled with disruptive technologies like containers, serverless computing, and microservice-based architectures. We also believe in high standards for quality with a zero-bug policy and zero downtime deployment approach.