Publicis Sapient

09.04.2022

Aleksandr Chernyshev
acme
chernyshev.alexander@gmail.com

Overview

Company VideoEX is an online video streaming company, similar to YouTube or Dailymotion. VideoEX is a global leader delivering innovative communications and technology solutions that improve the way customers live, communicate and work.

Goals

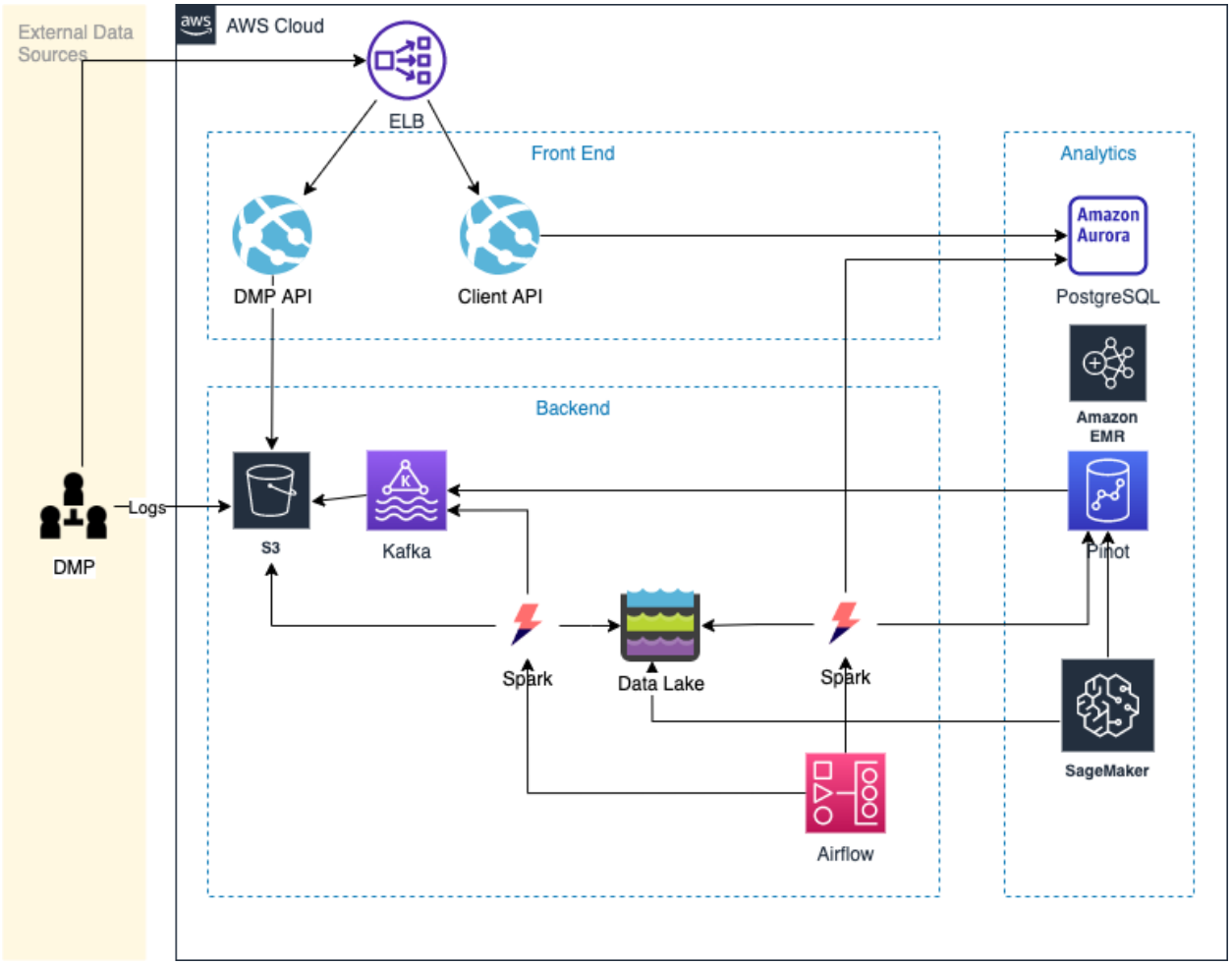
VideoEX's head of product has identified a problem of high drop-off rate for their home page i.e. users visiting the home-page and leaving the site without any action or watching any content. This is major problem for VideoEX as they have high

customer acquisition costs and want users to spend some time on their platform and watch videos in order to generate revenue. Their CMO wants to do a detailed analysis of the issue using existing dataset and requires dashboard of this data for decision making for future campaign spends.

Specifications

System design

The architecture is based on AWS cloud services. It could be migrated to any cloud. Following lambda pattern. In the spirit of minimalist.



External Data sources

Supposedly, that data comes to the system from the external data management platform to S3 storage.

Front End

Provides DMP API to ingest data into the system using Rest/GRPC protocols.
Provides Client API to interact with the clients.

Implementation: Scala ZIO-HTTP framework

Notes. Assumes a standard AWS infrastructure for web services with a minimum of 2 Availability Zones and Scale Groups behind an Elastic Load Balancer.

Backend

Confluent Kafka collects real-time data from S3 storage. Use Kafka KSQL component to build streaming applications to process data.

Spark is used to ingest data from Kafka into a Data Lake. Solves data cleaning and data quality problems. Spark Data format: Delta

Data Lake holds semi-structured data (bronze lake), and structured data (gold lake) on S3
Data lake format: Delta

Spark reads data from Data Lake (can use Delta live tables) to ingest data into Data Warehouse for analytics and Aurora PostgreSQL for client reports with limited data views.

Airflow - schedules spark jobs on AWS EMR.

Analytics

Apache Pinot - real-time analytics database. It has a Kafka connector to load incoming data from the topics. Provides a real-time view of data. Data in these tables have a small retention period as they need only for online analytics.

Notes: If Apache Pinot looks not very mature at this moment, replace it with Snowflake DB. To ingest data to Snowflake we already have Airflow cluster. You will be charged for Snowflake queries.

Spark ingests high-quality data from Data Lake to the domain-specific tables in Apache Pinot. This data is available for reporting and decision tools.

Spark ingests high-quality stripped data from Data Lake to the domain-specific tables to AWS Aurora PostgreSQL. This is reporting for the clients.
To support many clients, use master-slave replication to have a small latency.

Reporting

Tableau - Adhoc and Canned reports.

Machine Learning

Use SageMaker or/and Spark ML to train models on Data Lake sources, Pinot(Snowflake)

CI/CD tools

AWS Code Commit, Jenkins, MLFlow