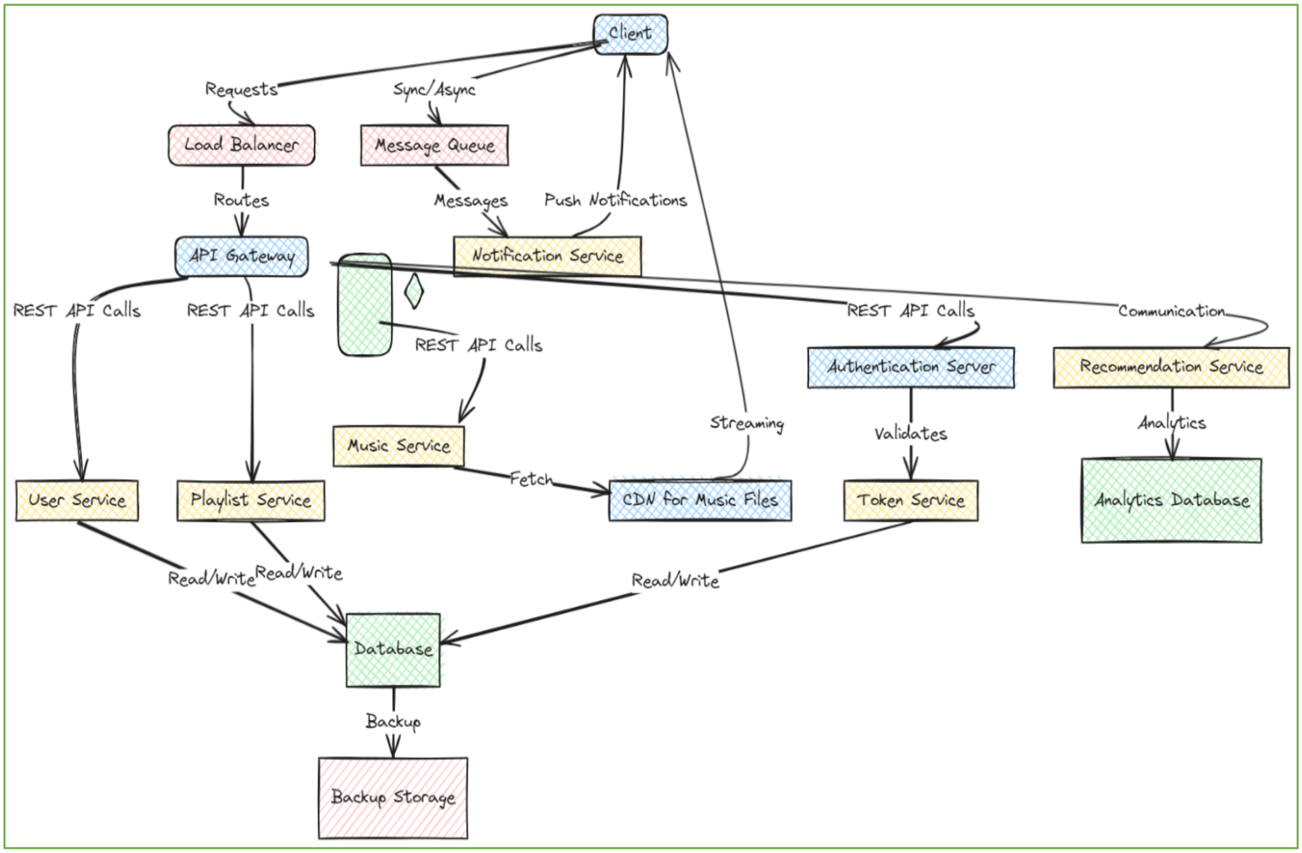
Spotify Architecture and Design Consideration

- **Prepared by Anupriya Sen(2022OG04013), Kumar Ashutosh(2022OG04019)**



Spotify End to End Architecture `

Design Consideration:

Designing a Spotify-like music platform involves several components. Here's a high-level overview of the tech stack and product choices. As there is a saying "*No design is good or Bad*" , here is our justification to suggest the below technical stack for the architecture design.

Frontend: React.js or Angular.js for building the user interface. Both are mature, widely-used JavaScript libraries with a strong community, which is a critical aspect when choosing a tech stack.

Backend: Node.js or Django can be used for building the backend. Node.js is a JavaScript runtime built on Chrome's V8 JavaScript engine, allowing developers to use the same language for frontend and backend. Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design.

Database: NoSQL databases like MongoDB or Cassandra can be used to store data. These databases provide horizontal scalability which is key for a large-scale music platform. MongoDB is a general-purpose, document-based, distributed database with scalability and flexibility. Cassandra is a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

Regarding trade-offs with respect to the **CAP theorem** (Consistency, Availability, Partition tolerance), in a distributed system, we can only choose two. For a music platform like Spotify, availability and partition tolerance are likely to be prioritized over consistency.

If our application design we want a flexible schema, ease of use, and rich querying, so MongoDB is a good choice for us.

During implementation, we have used Mongo DB cloud as the NO SQL Database to store the records.

however, if we are considering database that excels in distributed, high-volume, always-on environments, Cassandra may be more suitable choice.

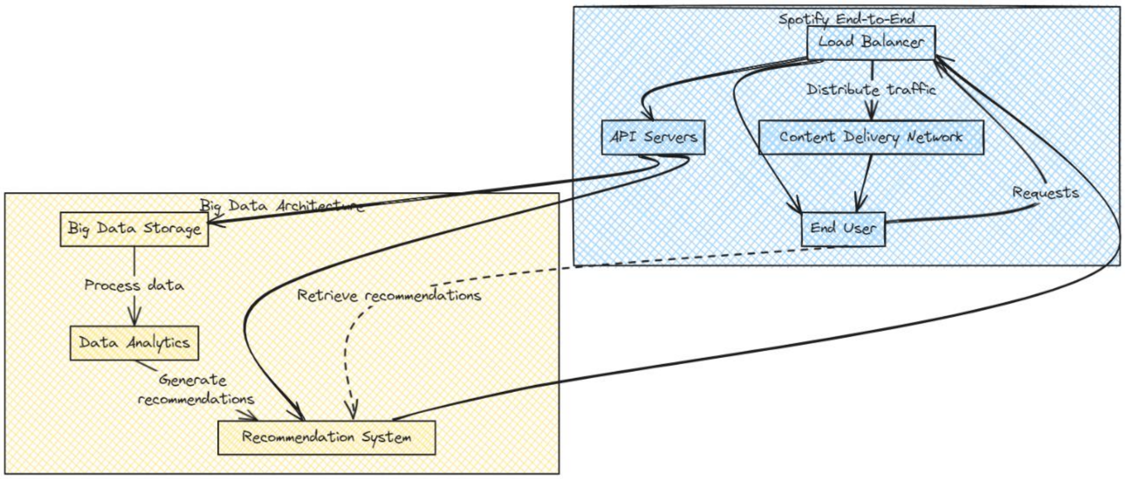
Load Balancer: Nginx or HAProxy can be used as load balancers. Load balancing refers to efficiently distributing incoming network traffic across a group of backend servers, also known as a server farm or server pool.

Batch Processing: Apache Spark can be used for batch processing of data.

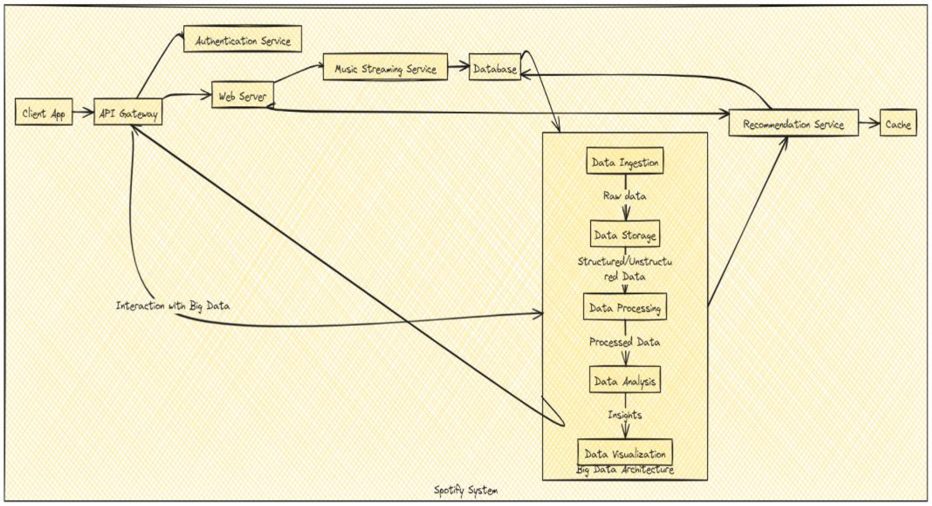
We have selected SparkSQL over Pig or Hive because of the following reasons -

- **In-Memory Processing:** SparkSQL, part of the Apache Spark ecosystem, leverages in-memory processing, making OLAP queries significantly faster compared to traditional Map-Reduce jobs.
- **Unified Processing Engine:** SparkSQL provides a unified platform for both batch and interactive processing, eliminating the need to switch between different tools like Pig and Hive, leading to a more streamlined and efficient architecture.
- **Optimized Query Optimization:** SparkSQL's Catalyst optimizer enhances query performance by generating optimized query plans, surpassing the optimization capabilities of Map-Reduce, Pig, or HiveQL.
- **Ease of Use:** SparkSQL simplifies complex data processing tasks with a SQL interface, enhancing usability and reducing the learning curve compared to traditional Map-Reduce or Pig/HiveQL approaches in a Spotify-like OLAP architecture.

Stream Processing: Apache Kafka can be used for real-time stream processing. Kafka is a distributed streaming platform that is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, and incredibly fast.



Recommendation System: For the recommendation system, a combination of collaborative filtering and content-based filtering algorithms can be used. Collaborative filtering algorithms suggest products to a user that similar users have liked. Content-based filtering recommends items by comparing the content of the items to a user profile.



Analytics Systems: Google Analytics or Mixpanel can be used for collecting and analyzing user behavior.

Data Analytics System: Apache Hadoop and Apache Spark can be used for data analytics. Apache Spark is an open-source, distributed computing system used for big data processing and analytics.