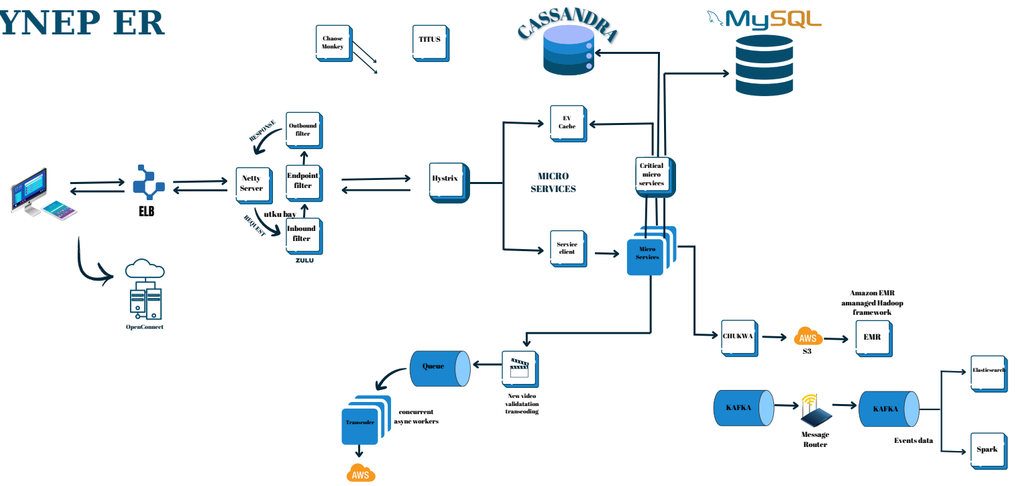
NETFLIX

ZEYNEP ER



- **WHAT IS NETFLIX'S ARCHITECTURE ?**
  -Netflix uses microservices architecture.

NETFLIX USES DIFFERENT DATA STORES COMPRISING BOTH SQL AND NOSQL FOR DIFFERENT PURPOSES.

- **WHY MYSQL ?**
  -MySQL, with its strong consistency and reliability, serves as the source of truth for most business data.
  -MySQL databases are used for managing movie titles, billing, and transaction purposes.
  -AWS EC2 Deployed MySQL is used to store the data.
  -The data is replicated across multiple data centers (cross-DC) to ensure high availability and disaster recovery.

- **WHY CASSANDRA(NOSQL) ?**
  -Netflix has a large user base globally, so it requires such DB to store user history.
  -It enables handling of large amounts of reading requests efficiently and optimises the latency for large read requests.

INITIALLY, THE VIEWING HISTORY WAS STORED IN CASSANDRA IN A SINGLE ROW. WHEN THE NUMBER OF USERS STARTED INCREASING ON NETFLIX THE ROW SIZES AS WELL AS THE OVERALL DATA SIZE INCREASED. THIS RESULTED IN HIGH STORAGE, MORE OPERATIONAL COST, AND SLOW PERFORMANCE OF THE APPLICATION. THE SOLUTION TO THIS PROBLEM WAS TO COMPRESS THE OLD ROWS...
NETFLIX DIVIDED THE DATA INTO TWO PARTS...

- **LIVE VIEWING HISTORY (LIVEVH):**
  -This section included the small number of recent viewing historical data of users with frequent updates. The data is frequently used for the ETL jobs and stored in uncompressed form.

- **COMPRESSED VIEWING HISTORY (COMPRESSEDVH):**
  -A large amount of older viewing records with rare updates is categorized in this section. The data is stored in a single column per row key, also in compressed form to reduce the storage footprint.

- **WHY WOULD NETFLIX USE BOTH SQL (MYSQL) AND NOSQL (CASSANDRA) DATABASES IN THEIR SYSTEM DESIGN?**
  -SQL and NoSQL databases serve different needs. MySQL, an SQL database, provides strong consistency and is great for transactional data, making it ideal for business operations. Cassandra, a NoSQL database, excels in scenarios that require high write performance and scalability, perfect for storing and processing high-volume data like user viewing history.

- **HOW IS SEARCH IMPLEMENTED IN NETFLIX ?**
  -Elasticsearch, with around 150 clusters and 3500 instances, provides real-time search and analysis capabilities to Netflix. It enables Netflix to derive actionable insights from a sea of data quickly.
  -Search is implemented using Elastic Search DB that enables users to search for movies, series by title, or any meta-data associated with the video. Elastic search provides the feature of full-text data search and ranking the data based on recommendations, reviews, rankings during search only.

- **SPARK FOR RECOMMENDATION SYSTEMS AND MORE**
  -Netflix makes extensive use of Apache Spark for a variety of tasks, but its most well-known use is for powering Netflix's recommendation engine. This engine uses both collaborative filtering and content-based filtering to curate a personalized set of recommended movies and shows for each user. The process is complex and computationally intensive, making Spark's distributed computing capabilities a perfect fit.
  -The recommendation engine also takes care of choosing the right artwork or thumbnail for each movie or show. The system generates multiple thumbnails and uses advanced machine learning techniques to select the one that is most likely to attract a particular user based on their viewing history.

- **KAFKA AND CHUKWA FOR EVENT LOGGING**
  -Netflix generates about 500 billion events per day from various sources, including view activity, error logs, and performance events. Kafka, a highly scalable and fault-tolerant messaging system, serves as the first point of contact for these events.
  -Chukwa, a data collection system built on top of Hadoop Distributed File System (HDFS), transports the events to a centralized store like AWS S3 for long-term storage and analysis.

- **AWS SERVICES – ROUTE53, SNS, S3, AND MORE**
  -Finally, it's important to note that Netflix extensively uses various Amazon Web Services (AWS) to supplement its own infrastructure. AWS Route 53 is used for DNS, Amazon Simple Notification Service (SNS) is used for push notifications, and Amazon S3 is used for storing massive amounts of data, including video files and user viewing history.