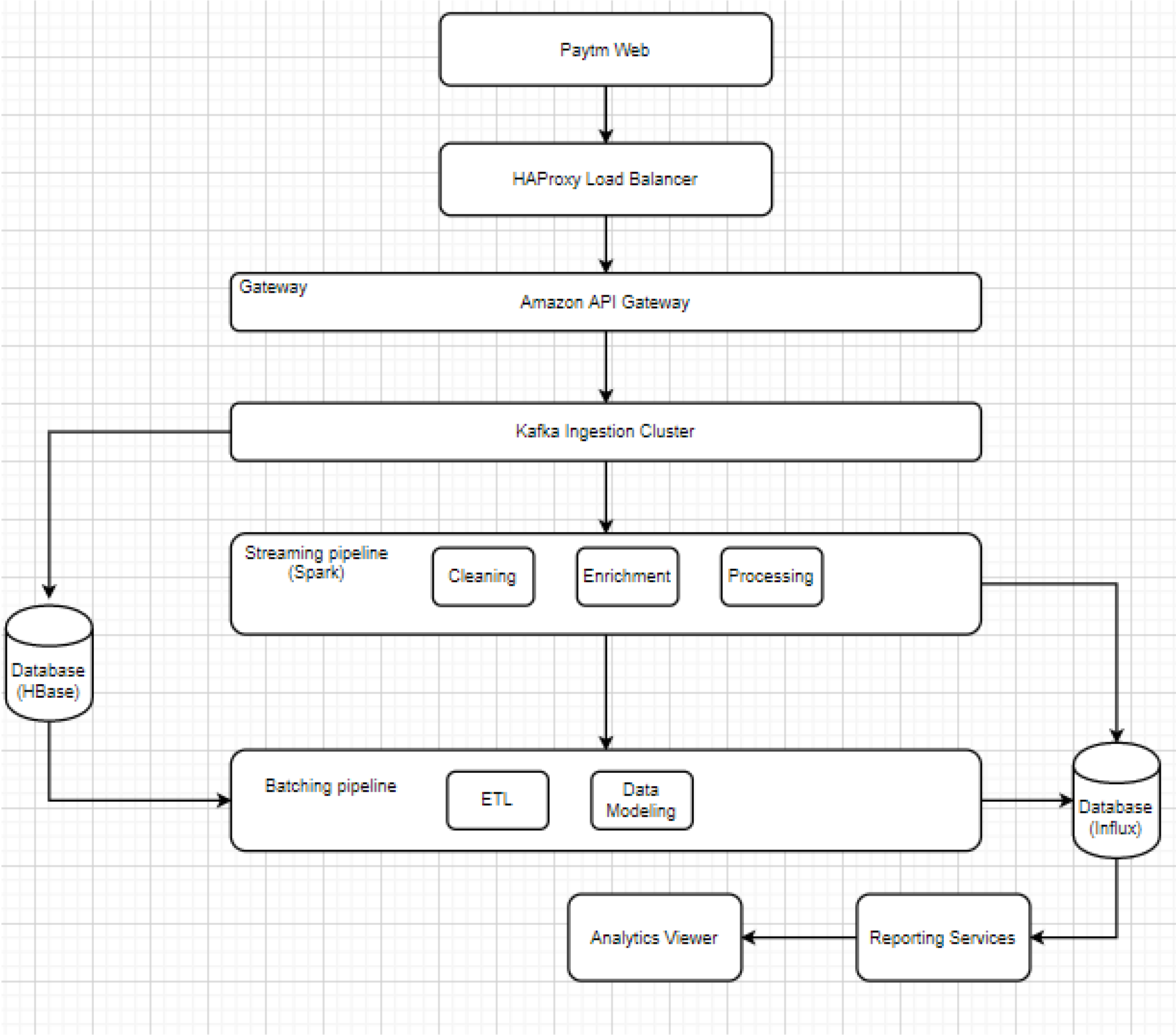
Question

https://github.com/PaytmLabs/SDEChallenge

Architecture and key systems

The whole backend system is adopting a micro services-oriented architecture

- Load Balancer: HAProxy
- Monitoring: ELK and Dynatrace
- Data Pipeline: Amazon API Gateway, Kafka, Spark
- Cluster Management: Kubernetes
- Workflow Management: Airflow
- Data Warehouse: HBase
- Time Series DB: InfluxDB
- Reporting Services: REST Microservices via Spring Boot



High Availability and Scalability

Using Kubernetes enable us to manage high available clusters with minimum downtime and can do deployment/node recovery. So, this is best to support billions of request every day

Scalability and managing cluster dynamically is one of the key features of Kubernetes. Using Kubernetes will help to enable pods/container increase for future independently.

System Implementation

HAProxy Load Balancer

HAProxy, which stands for High Availability Proxy, is a popular open source software TCP/HTTP Load Balancer and proxying solution.
It will be used to improve the performance and reliability of a server environment by distributing the workload across multiple servers.

Async Data Emitter

Data emitter must be integrated into Paytm web apps and mobile apps, which will send events to data collecting layer using amazon API gateway. It will load an async script that assigns a tracking cookie to a user if it isn’t set yet. It will also sends an XHR for every user interaction, like a page load. These XHR requests are then processed and raw event data is stored and scheduled for aggregation processing. Depending on the total amount of incoming requests, the data will also be sampled.

Event Schema

Unified event format is essential:
1) Unified event structure.
2) Using **json** as transfer protocol

Data Ingestion

Apache Kafka is used for building real-time streaming data pipelines that reliably get data between many independent systems or applications.
In this solution it will allow:
· Publishing and subscribing to streams of records
· Storing streams of records in a fault-tolerant, durable way
- Spanning multiple datacenters spread across geographies.

It also will provide a unified, high-throughput, low-latency, horizontally scalable platform being a reliable and mature messaging system.

Data Processing

LAMBDA structure is being implemented using below technologies

Spark: Apache Spark is an open source and flexible in-memory framework which serves as an alternative to map-reduce for handling batch, real-time analytics, and data processing workloads. Here it will be used for improving the processing time.
Kafka: Kafka is used for internal communication between the different streaming jobs.
HBase: Apache HBase is an Open source distributed column-oriented NoSQL database that runs on top of Hadoop Distributed File System (HDFS). It is natively integrated with the Hadoop ecosystem and is designed to provide quick random access to huge amounts of structured data. HBase in this solution can be used to to store the processed data.

1) **Spark streaming pipeline** : For generating near real-time reporting results.
2) **Batch pipeline** : It is used for large event processing which serves as a complementary/correction of streaming **Airflow** pipeline and also serving other underlying BI/data science tasks.

Reporting Layer

Using a microservices approach to application development can improve resilience and expedite the time. With fine-grained services and lightweight protocols, microservices will offer increased modularity, making applications easier to develop, test, deploy, and, more importantly, change and maintain.
With microservices, the code is broken into independent services that run as separate processes to process data reports/trends to end users. This layer will also have support for creating and deploying reports for better data analysis

Monitoring Technologies

Monitoring is one of the key features and below tools need to be implemented
- **ELK Monitoring:** Onboard ELK for log monitoring all services and check for issues
- Integrate with datadog to setup alerts.
- **Dynatrace**: Check overall system health and perform validation via **Dynatrace** setup.

Dynatrace is a software intelligence company providing application performance management (APM), artificial intelligence for operations (AIOps), cloud infrastructure monitoring, and digital experience management (DEM).