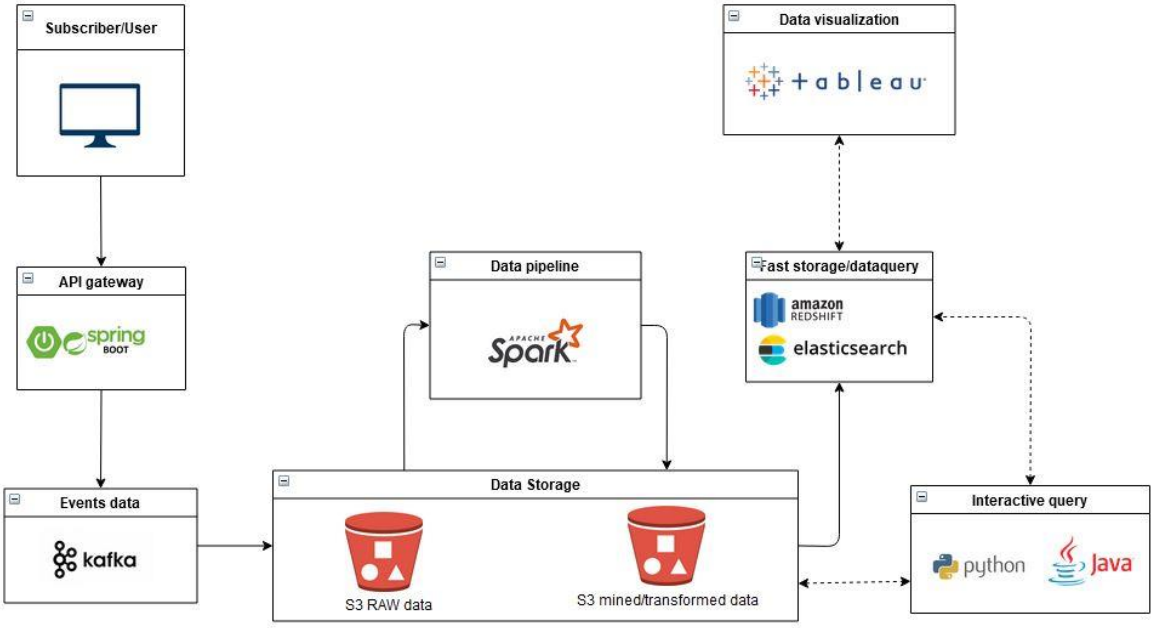
Design Question

Design A Google Analytic like Backend System. We need to provide Google Analytic like services to our customers. Please provide a high level solution design for the backend system. Feel free to choose any open source tools as you want.



The Above design will handle all specified requirements for large amount of data for billions of events. Large query reads, minimum downtime and the ability to interact with the data for reprocessing. Below is a walkthrough of the above design and a bit of a summary of each component.

Subscriber/User

The design is made to handle user event data gotten from the end user interfacing with the application, this application is registered and sends event data that are needed for the analytics, it can be the any sort of application needed for this process(website, app, email, etc). When an event needed for analytics is triggered, data is sent to the API gateway.

API Gateway.

The API Gateway servers as the door to all other microservices and also serves as a level of security. The API gateway does not manipulate the raw data being sent but does some authorization and authentication with the interfacing subscriber. The events data is further pushed to a Data collection and streaming tool, in this case kafka, and is arranged based on required topics.(Other microservices may be between kafka and API gateway for separating the API gateway security functionality).

Event Data Collection

In this design, Kafka is used as it provides high throughput, fault tolerant, real time data handling. Kafka is able to process about millions messages per second and Trillions of data per day. The events messages in this case are streamed based on topics and those topics can be created based on the category of events and how the raw data wants to be analysed. This data is then passed on to a data storage which is this case is S3 buckets.

Data Storage

S3 was chosen in this case because it is known to provide highly-scalable object storage, easy to use, store and retrieve any amount of data. It can store files of up to 5TB and most importantly known for is 99.99% uptime. S3 buckets can be created to store raw event data gotten from the users and then can be further passed through a data pipeline for mining. There are two categories of S3 buckets in this design, on for raw data and the other of transformed data. The raw data is passed trough a data pipeline and transformed, this transformed data is put into the S3 category of buckets holding transformed data. The mining/transformation process is carried out using apache spark.

Data Pipeline

Apache spark is used in this design to transform the data based on what is required from the raw data and to provide better insight of the raw data. This makes the data more useable to for analytics. Spark can process billions of data per second and it is appropriate for this job.

Fast Storage/data query

After the data has been sent to the S3 bucket containing the transformed data, a portion of the data needed for analytics can be send to a fast storage and query service or tool like amazon redshift or elastic search. These tools provide the ability to query large amount of data seamlessly. They are the appropriate tools to carry out million of queries on the transformed data without any issues.

Interactive Query

In order to reprocess historical data and an view every data thoroughly, interactive queries can be made using languages like Java or python for these interactions. These interactions can be done to the data storage services and the fast storage service.

Data visualization

Finally to visualize the data, a tool like Tableau can be used for this.