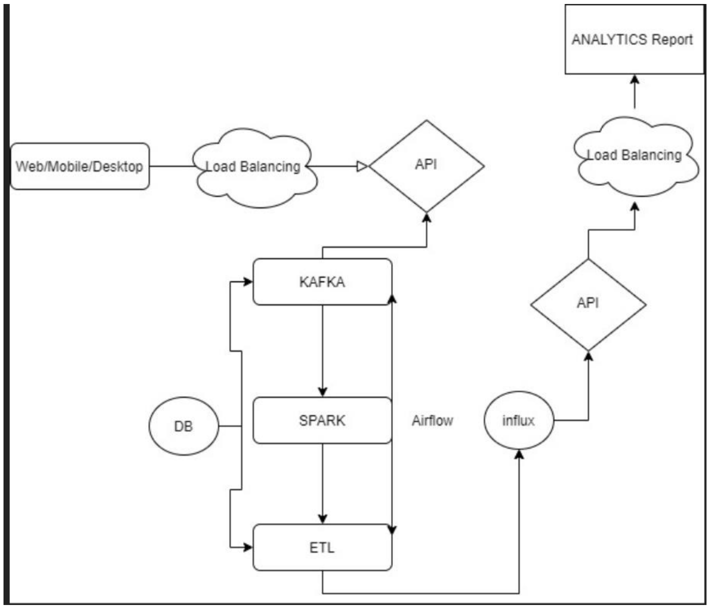

<u>GA-Google Analytics</u><u>: </u>Every visitor is basically tracked by GA code that sends a request every time user clicks on a button that basically generate events (clicked on a image, footer, header, anchor links,buttons, etc) that are then sent to the server.

<u>Apache Kafka: </u> Publish and subscribe to streams of records, like a message queue, Process streams of records as they occur, building real-time streaming applications that transform or react to the streams of data

<u>Apache Spark</u> is open source, general-purpose distributed computing engine **used for** processing and analyzing a large amount of data

1. Now, on every interaction, /POST request will be made to our server with all the event data which will be further PUBLISHED to our Apache Kafka clusters (topics with the payload) .

   A typical Kafka broker instance can handle hundreds of thousands of reads and writes per second.

2. We can use Kafka's Producer API to publish a stream of records, Consumer API to subscribe to different topics, Stream's *API*, consuming an input stream from one or more topics and producing an output stream to one or more output topics.
3. Zookeeper can be used on Kubernetes for effectively managing the Kafka clusters.

4. Now, once we get Direct Streams from Kakfa, we can use Apache Spark to perform a series of operations, time-series manipulations using Spark-TS library

5. Now, we need to have a Data Pipeline, this can be automated or "coded" in **Airflow** so we can catch exceptions and instantiate pipelines programmatically.
6. We then can use Airflow's Connections to store data in our warehouse and further we can use something like **influx db** to pull some analytics reports to show to the client.