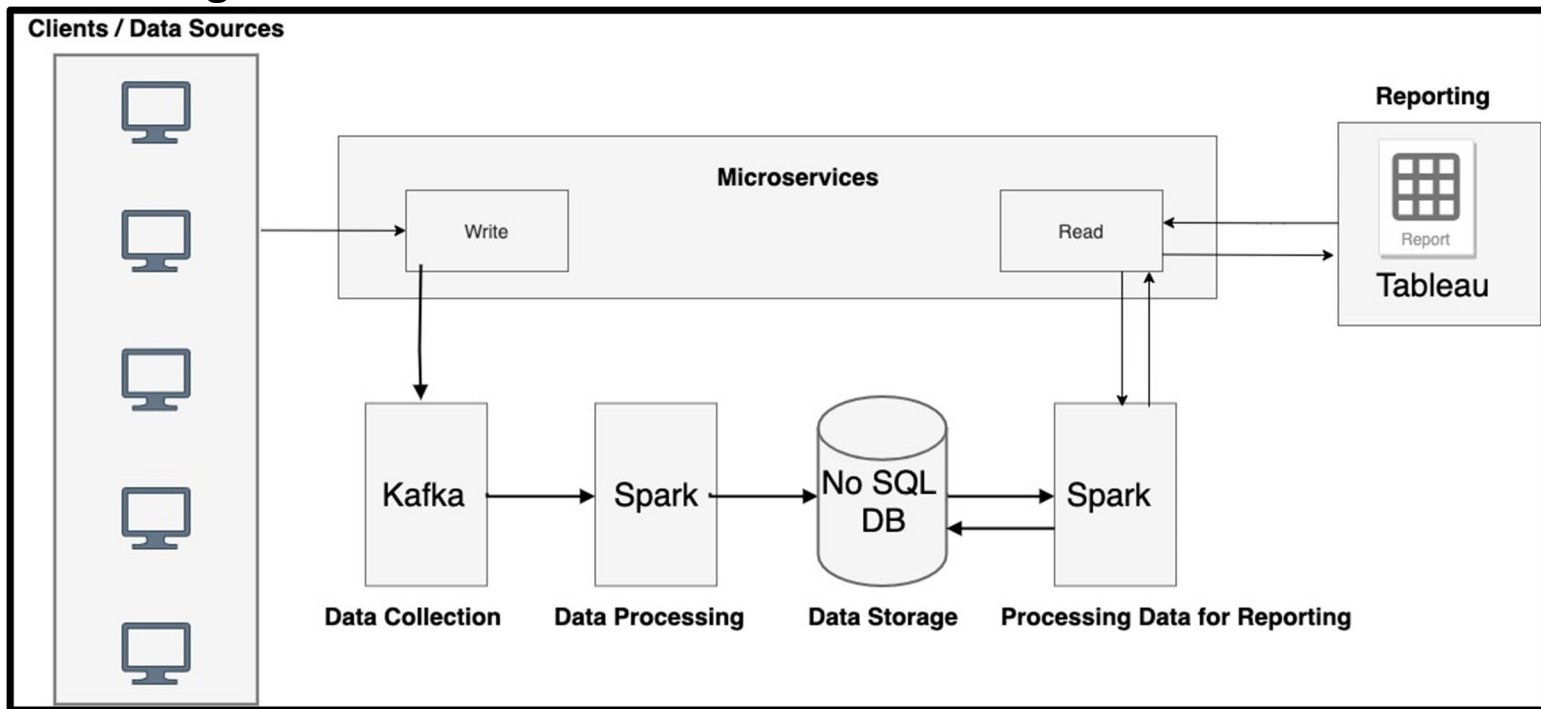


# High Level Design For Google Analytics like Backend System

## I. Design :



## II. Overview :

This system has the capability to act like google analytics backend engine.

To start with, we need to have a javascript which will be embedded on the client side so that it can capture the analytics data. This would send the analytics data to our 'Write' microservice which in turn will send the data to Kafka. This data would be sent on specific topics that must be created prior to this. The data now needs to be processed so Spark would consume these events and would further transform and process these. Once this is done, we need to persist the data in the DB because user can ask for reports for any past data.

Now, a user would also be required to build the reports.

Tableau would be used for visualization. With the use of Connectors, Tableau would hit the 'Read' microservice which would again use spark to process the queries faster and to ease the processing. This would in turn get the data from the No Sql DB in which the data has been persisted. The reports will then be generated.

So, in this way, this system would behave same as Google Analytics Backend Engine.

## III. Detailed Description :

Product	Details
Clients/DataSources	<ul style="list-style-type: none"> <li>• A javascript needs to be embedded on the client</li> <li>• This script will send the analytics data to our service</li> <li>• This would be asynchronous so that it does not impact the performance on the client side</li> </ul>
Write Micro Service	<ul style="list-style-type: none"> <li>• This microservice will be used to write the events data to Kafka</li> <li>• This will work as abstraction layer where we will have the capability to make changes in the underlying structure later on (if required)</li> <li>• This will only push the data to Kafka</li> </ul>
Kafka	<ul style="list-style-type: none"> <li>• Kafka is reliable, durable and has a flexible publish-subscribe model that scales well</li> <li>• It will be used to capture the events which will be used further for processing</li> <li>• Different topics will be created and events will be published on these topics</li> <li>• There would be subscribers who would be listening to these topics and will consume the events</li> </ul>
Spark	<ul style="list-style-type: none"> <li>• Spark will be used for processing the data and transforming it</li> </ul>
No SQL DB	<ul style="list-style-type: none"> <li>• The data is then pushed to a No SQL DB so that we can persist the data and the users can query any data at any point of time</li> <li>• So, in case any bug occurs with the reporting logic, we will always have the data and new reports can be generated anytime</li> <li>• Cassandra is a good choice as it stores the data column wise and also scales well and is fault tolerant.</li> </ul>
Spark	<ul style="list-style-type: none"> <li>• This will help in processing the data as per the queries as users can query any type of data</li> <li>• This will help in faster queries to be used in reporting</li> <li>• This will also help in reprocessing</li> </ul>
Read MicroService	<ul style="list-style-type: none"> <li>• This microservice would be used by Tableau to fetch the data from the No Sql DB.</li> <li>• We can use Web connectors in Tableau so that it can consume the response of this service</li> </ul>
Tableau	<ul style="list-style-type: none"> <li>• This will be used as a virtualization tool to build up the reports</li> </ul>