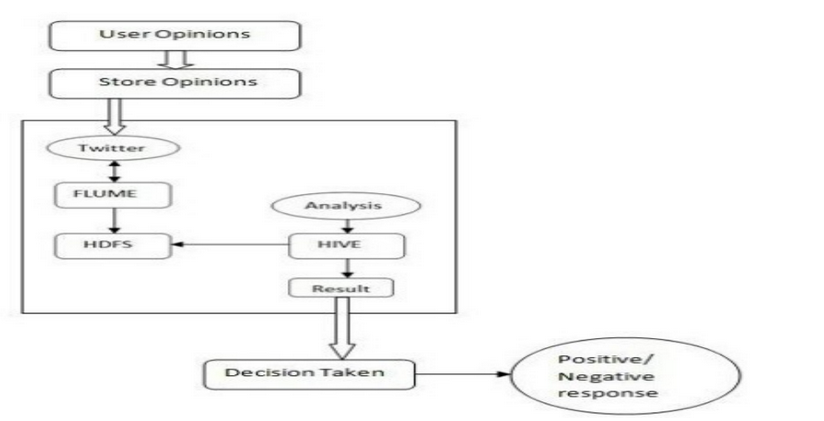
# Methodology

A tool that can easily handle the data in minimum time limit is required to analyze large and complex data.for this purpose Apache Flume and Hive query language were used. These are open source, scalable and fault tolerant framework.

Data analysis transformation for large and complex data was performed using HQL.HQL is similar to SQL i.e Structured query language.



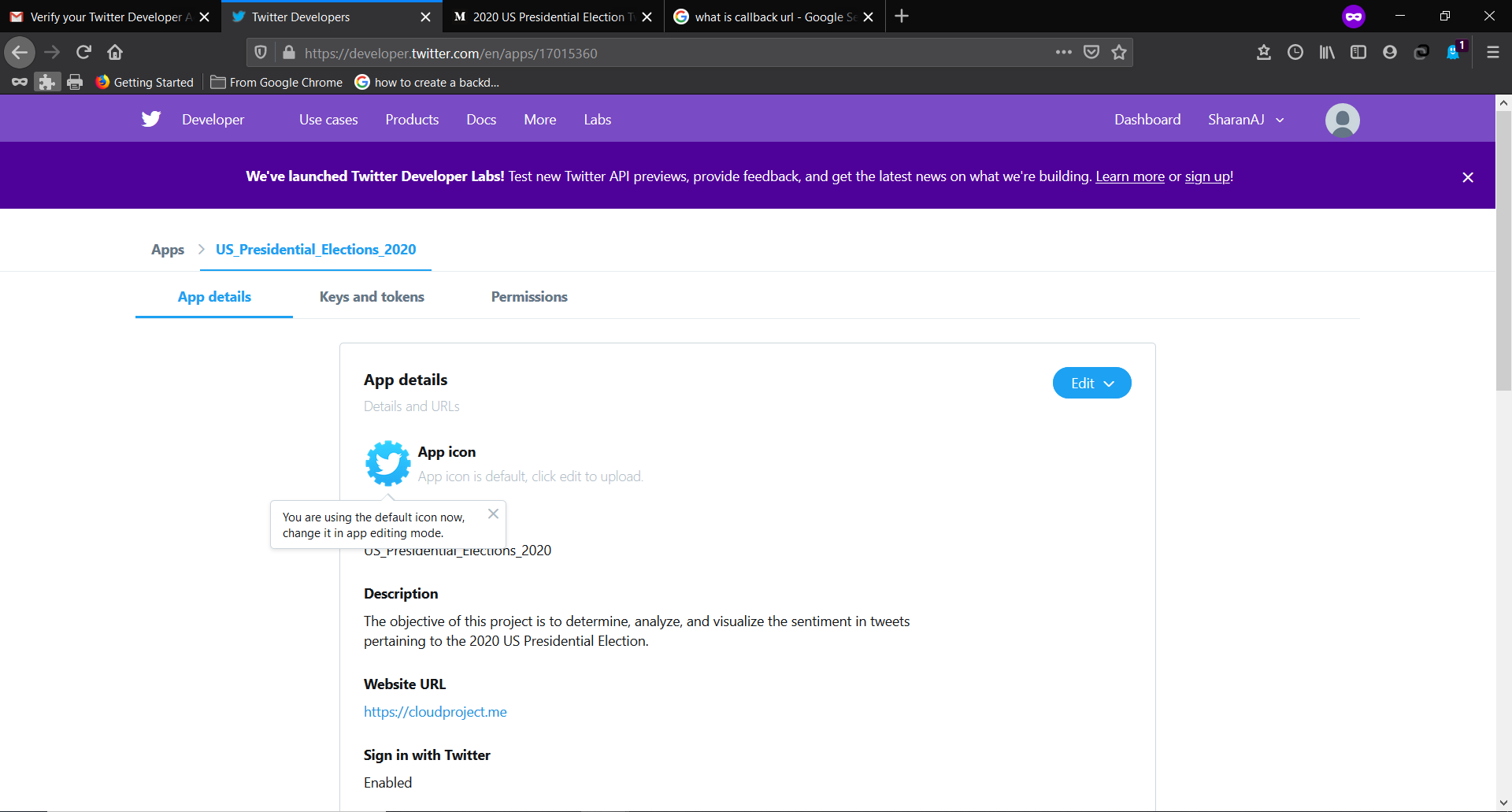
The various steps required for data ingestion, data storage and data process are:

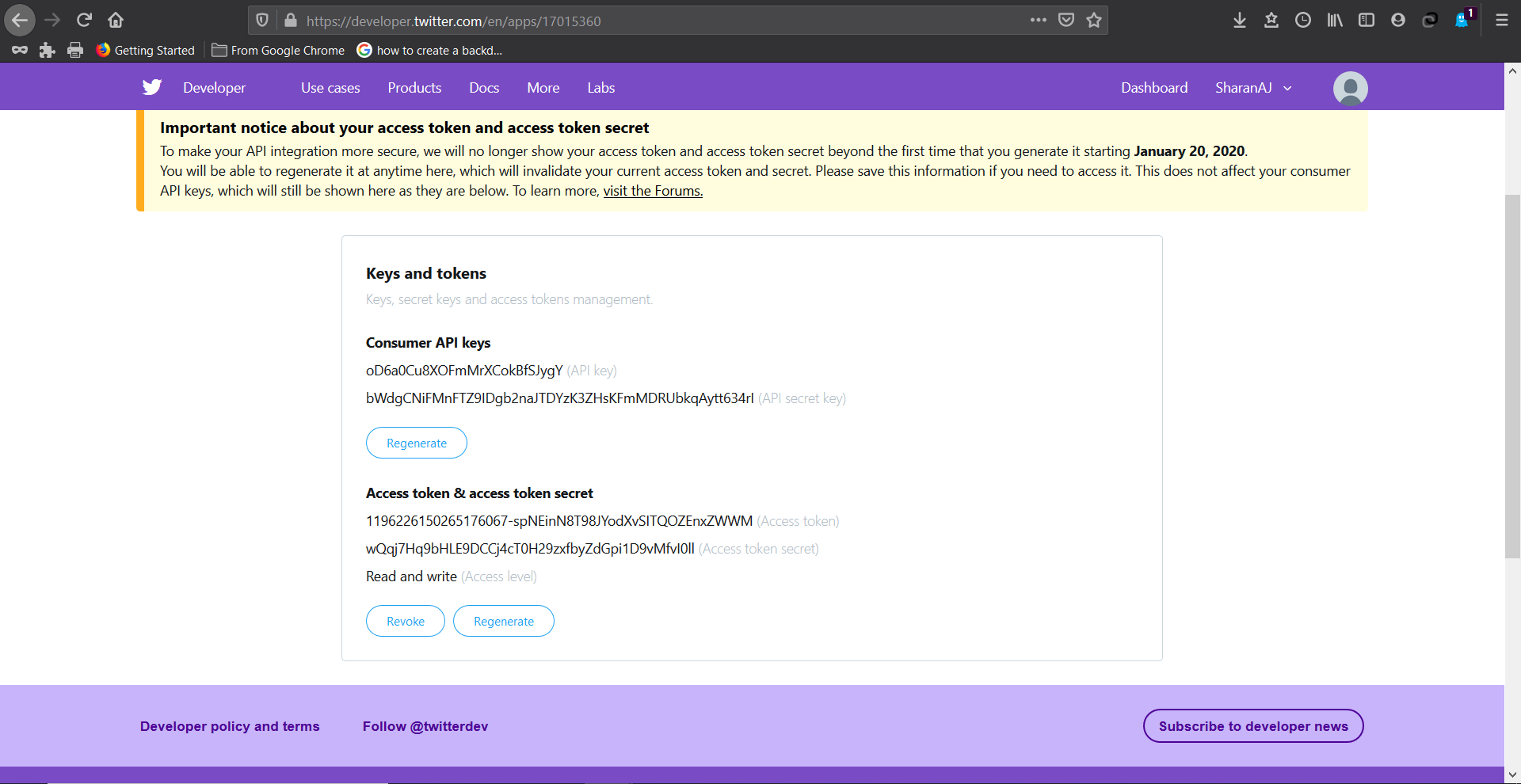
1. Create a Twitter Application
2. Data Ingestion using Apache Flume
3. Create an external table using Hive

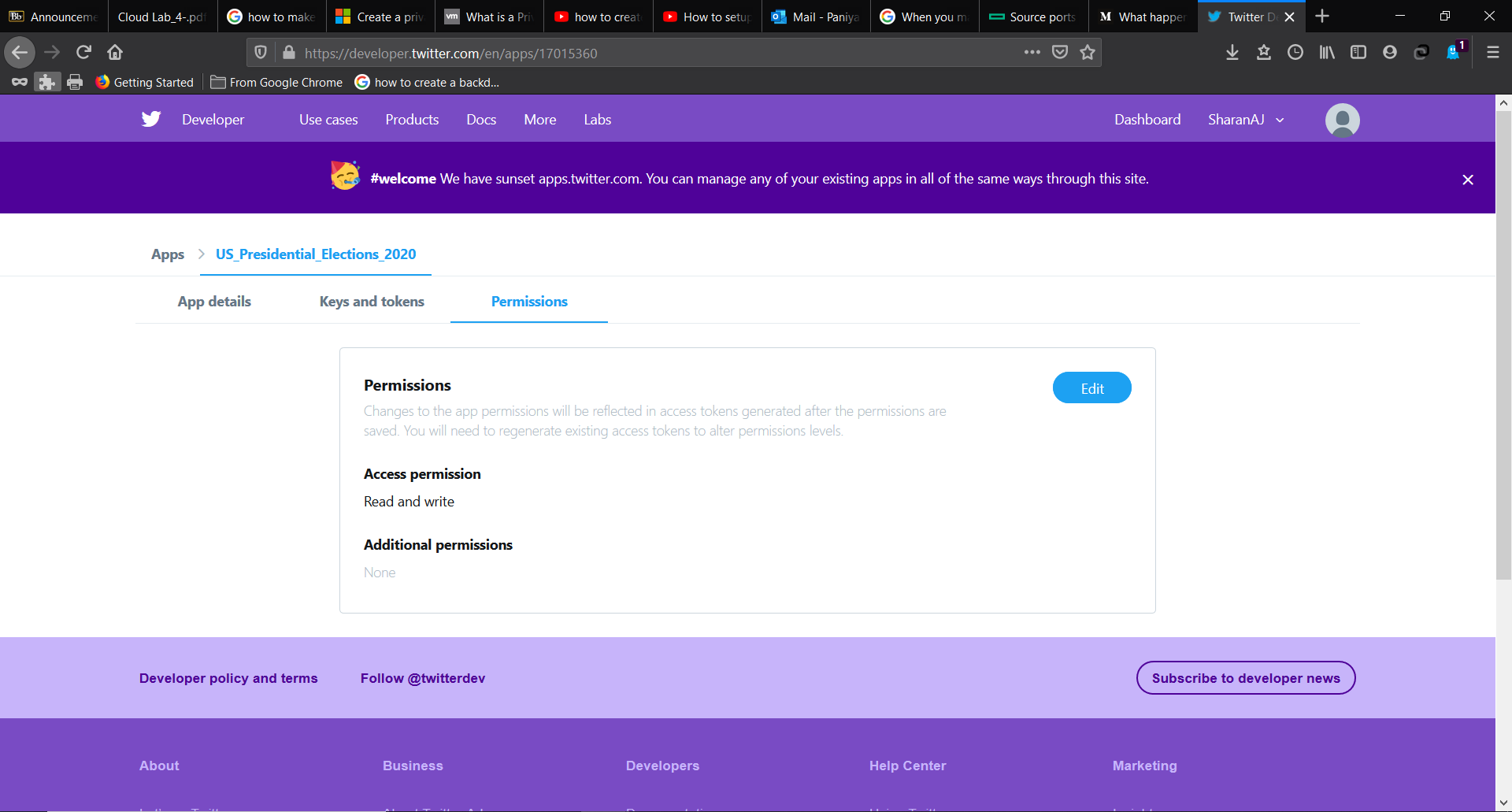
**Creating a Twitter Application**

In order to ingest the data, a twitter application was created along with the required consumer keys and access tokens.

* The twitter application was created using a developer twitter account by logging into developer.twitter.com/app website. The appropriate username and password credentials were used to log in.
* All the required details were provided after which permissions were obtained from twitter to use their data.
* Once the twitter application was created the following page appears which contains the application details such as description and website URL.



* The respective Consumer API keys such as API key and API secret key, Access tokens and Secret Access tokens were generated as shown in the below image:
* Once the keys and tokens were obtained the permissions were also set as shown below:



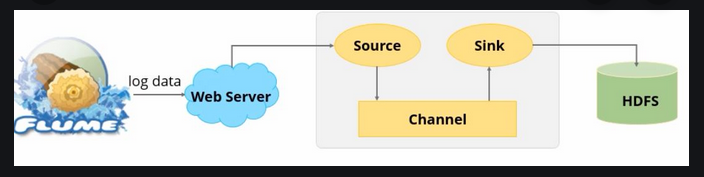
All the above details that were obtained during the twitter application creation was needed to be provided in the flume configuration file which was used further for the data streaming.

**Fetching twitter data using Apache Flume**

**Apache Flume**

Apache Flume is a toll/service/data ingestion mechanism used for collecting, aggregating and transporting large amounts of streaming data such as log data, events from various webservers to a centralized data store.

Its is a highly reliable, distributed, and configurable tool that is principally designed to transfer streaming data from various sources to HDFS.



The architecture of flume tool is simple. It contains three important roles.

**Source:** It is responsible for the data coming into the queue/file.

**Channel:** It is responsible for the connection between source and sink.

**Sink:** It is responsible for data flow of the queue/file.

**Log Data:** The data that will be coming from various data sources like web servers, cloud servers, enterprise services need to be analyzed. The produced data will be in the form of log file and events. The data in the log file known as log data.

After the application was created in the twitter developer site, data is fetched rom twitter. The following steps were to be followed.

* The flume is configured to the twitter application by using consumer key, secret key and access token generated in the twitter application.

.