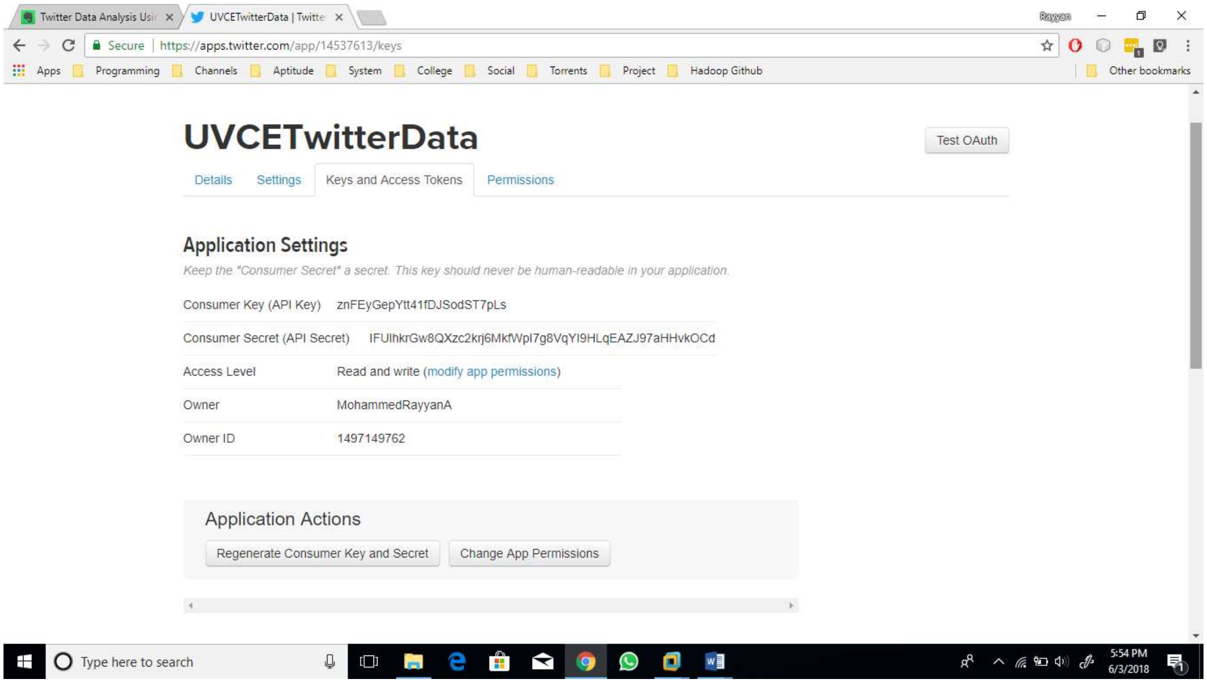
# CHAPTER 4

# SYSTEM ARCHITECTURE

## 4.1 Twitter Data

Twitter Developer Account can be created at Twitter Developers apps Page. In this page we need to provide valid twitter account page in the website field from which we need to get streaming data. If we provide valid details on this page we will get our app created as shown in below screen shots.



UVCETwitterData

Application Settings

*Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.*

| | |
|---|---|
| Consumer Key (API Key) | znFEyGepYtt41fDJSodST7pLs |
| Consumer Secret (API Secret) | IFUlhkrGw8QXzc2krj6MkfWpl7g8VqYI9HLqEAZJ97aHHvkOCd |
| Access Level | Read and write (modify app permissions) |
| Owner | MohammedRayyanA |
| Owner ID | 1497149762 |

## 4.2 Mathematical Model

Our proposed system is analysing system which is based on the mechanism that analyses User Tweets using Hashtags and Keywords. The proposed system collects tweets using this Hashtag's which are nothing but the popular personalities/Parties. General public orientation toward these parties can be studied using the tweets the people have posted on the

Tweeter. Tweeter is generally lauded by academicians, journalists and Politicians; for its potential political value. Many politicians make use of this micro blogging site to express themselves in the limit of 140 characters. These tweets can be categorized on various policies such as geo-location analysis to analyses the peoples view for that particular area which might help parties to design their winning strategy. The proposed system mainly focus on collection of tweets to make volume analysis to and out the popular days of election ; A trend analysis to and a popular or trending party/candidate and a sentiment analysis to actually bifurcate the positive and negative tweets for the party/candidate so that making trend analysis on this tweets can help this party/candidate to act accordingly to improve their reputation at the same time it might help user to actually make a clear opinion about any party/candidate. This will be conducted in 3 phases. To brief about it the phase one is connecting with tweeter and downloading the tweets. The second phase deals with loading these tweets on HDFS for further analysis and the third phase is the actual analysis and they are volume analysis, Trend Analysis and Sentiment analysis.

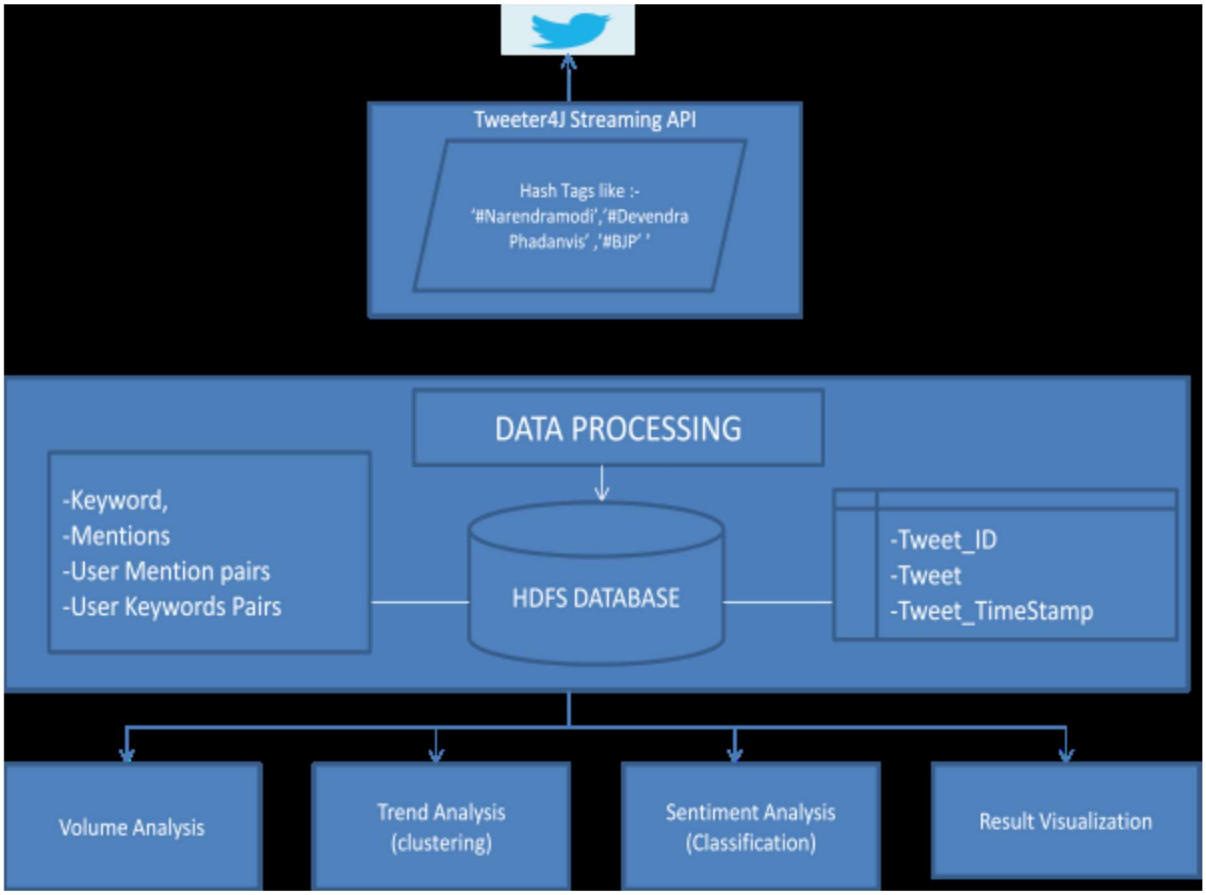


**Fig 4.1 Mathematical Model**

P= {I, O, F, U}

Where

I: Input

O: Output

F: Functions

U: User

Where

I= {U, UT, HL}

Where

U = User having a tweeter developer profile

UT = User Tweets which will be downloaded using authentication key

HL = Hashtags list provided as input for downloading election related tweets

O = {UP, PT, TA, VA, SA}

Where below are the output generated from system processing

UP= Retrieved User Profile details

PT= Processed Tweets by removing unwanted keywords stop words etc.

TA= Trend analysis on processed tweets will provide trending topic/Politicians

VA= Volume Analysis will provide the analysis based on date wise or location wise volume analysis

SA= Sentiment Analysis of tweets by categorizing them in positive, negative or neutral sentiments

U = {SV, TU, A}

Where

SU = System Visitor

TU = Tweeter User whose tweets are used in system as input

A= Administrator

F= {F1, F2, F3, F4, F5}

Where

1) Function F1: This function download the tweets through secure authentication using OAuth

2) Function F2: This function process the downloaded tweets for stop word removal

3) Function F3: This function performs volume analysis using Map Reduce Framework

4) Function F4: This function performs trend analysis using clustering of user, parties and related details.

5) Function F5: This function performs sentiment analysis by classifying tweets into positive, negative and neutral types

## 4.3 Software Environment

### 4.3.1 Hadoop Framework

Apache Hadoop is good choice for twitter analysis as it works for distributed big data. Apache Hadoop is an open source software framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the

MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. Hadoop framework includes different modules like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase for different functionality as shown in below diagram. I will be using FLUME and HIVE for twitter analysis.

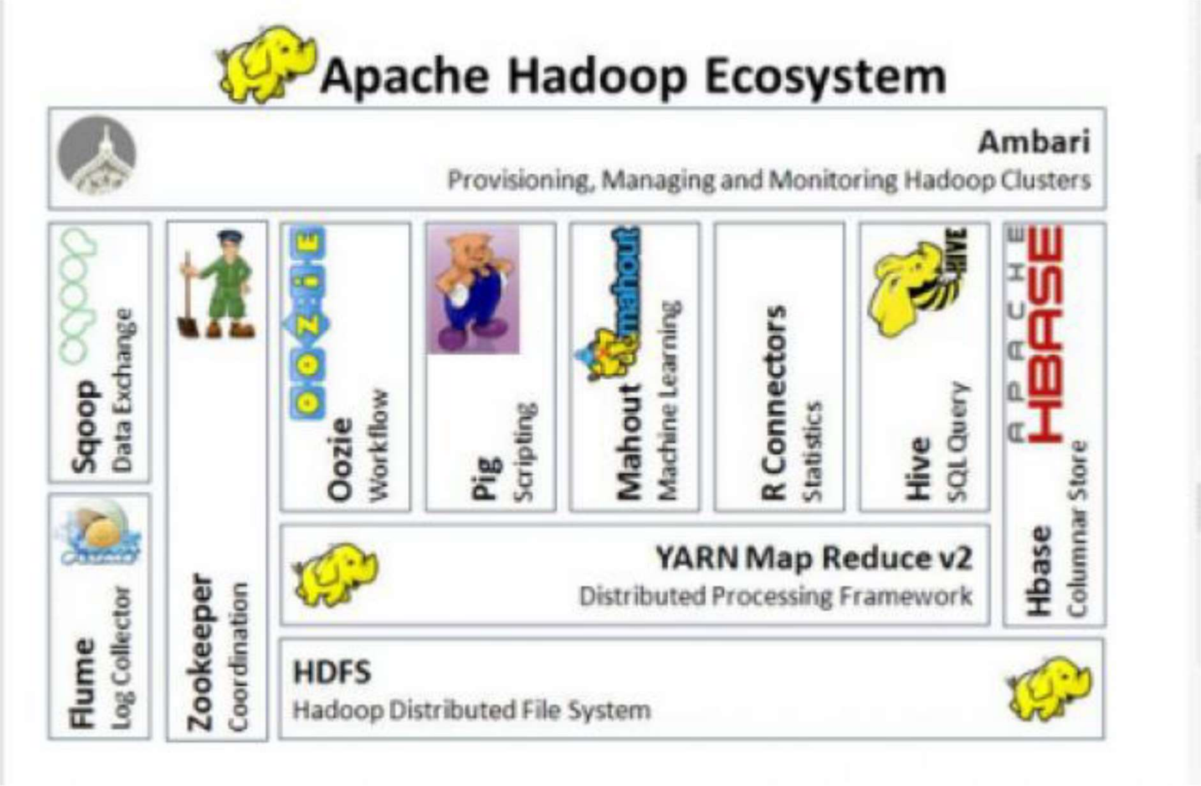


**Fig 4.2: Apache Hadoop Framework**

### 4.3.2 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is designed based on the Google File System (GFS), the master/slave architecture. The master consists of a single name node and one or more data nodes. The name node manages the metadata and data nodes stores the actual data. The name node decides the mapping of blocks to the data nodes. The data nodes take care of both read and write operation in the file system.

Hadoop use HDFS (Hadoop Distributed File System) file system. The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data. Benefit of using Hadoop is distributed storage, Distributed Processing, Security, Reliability, Speed, Efficiency, Availability, Scalability and lots more. This is the reason of using Hadoop for tweet processing.

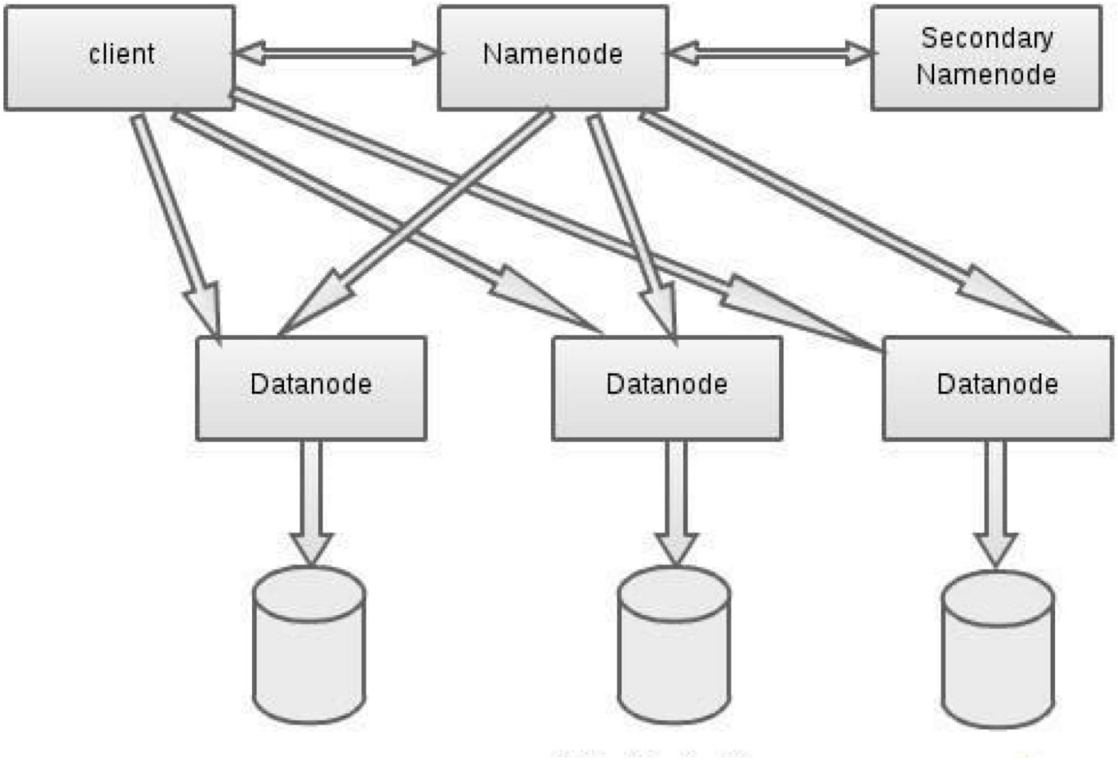


**Fig 4.3: Data Replication**

### 4.3.3 Hadoop Map Reduce

Hadoop map reduce is a software designed as parallel architecture running on large clusters. The map reduce algorithm consists of two important tasks, map and reduce task. The map task splits the data set into tuples (key/value pairs) and sorts the output and is fed as an input to the reduce task. The input and the output of the job are stored in file system. The process of scheduling, the execution of the failed jobs, monitoring is carried out using the framework.

Map reduce are mainly designed for the processing of large amount of data sets in the computer system in real-time technology. Execution engine Map reduce is consist of two functions map 0 and Reduce O. In this paper, we are study about the map reduce function. Big data doesn't have easy processing the in complex data sets or large sized data. User can make their own logic in map reduce programming model by already define a customized map () and reduce () functionality. First we Creates a list of intermediate data pairs with the help of map function takes an input data pair. Now in map present run time pairs or group together and all intermediate data pairs it is based on intermediate data. Now the job must be store in the file system with the help of input and output of the job.

Generally the large amount sized data sets data processing is very hard to operate. We can say it is very difficult to process the data. Managing thousands of record file and store the history data for the future development and more than thousands of processors to manage a parallelization and distribution file system to make environment. It is not easy to analyze and store the data without using

HADOOP. Now in this situation map reduce is available to provide a best solution of this type of issues, and also it is provide to solid solution for the distributed and parallel input output scheduling.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

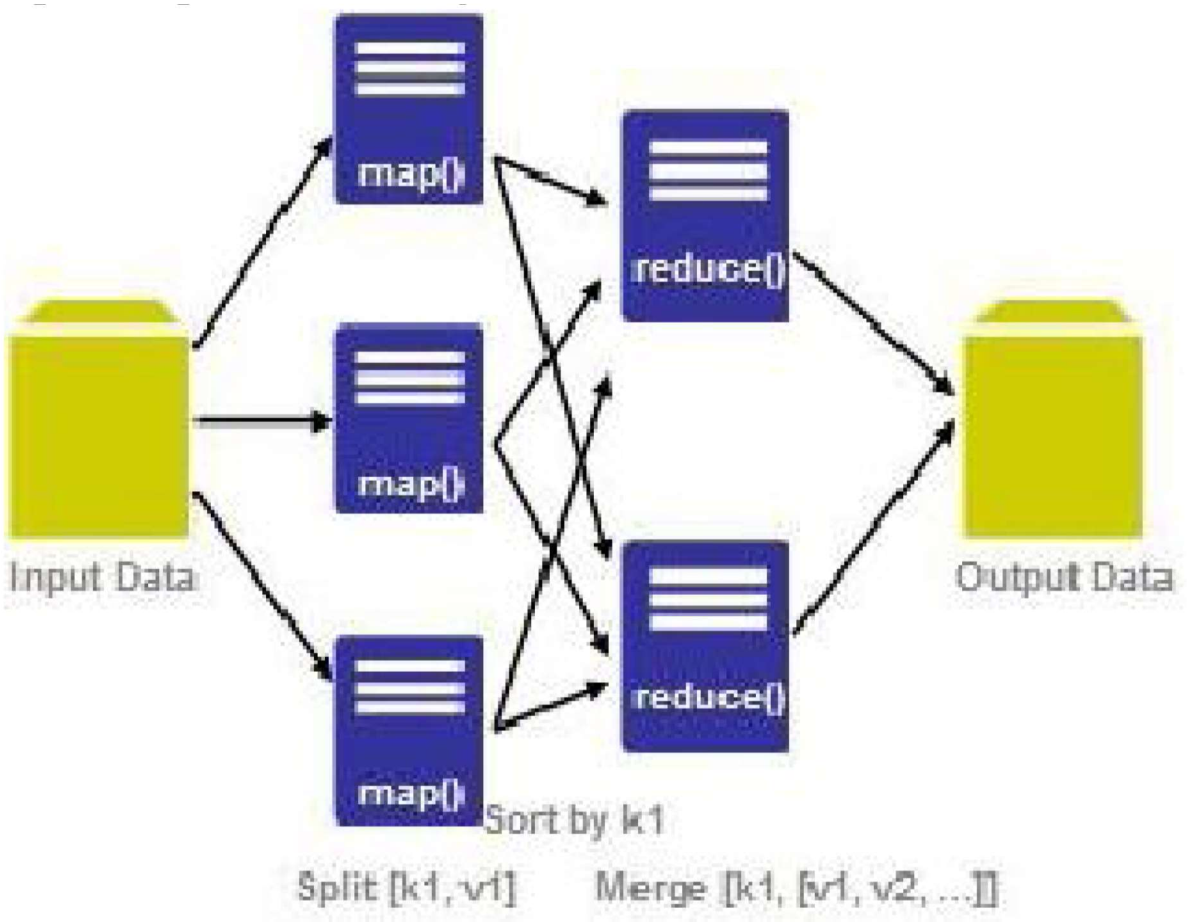


**Fig 3.3: Map Reduce**

### 4.3.4 Yarn

YARN (Yet another Resource Negotiator) is a cluster management technology, in the second-generation of Hadoop, designed from the experience gained from the First generation. YARN provides a central platform to deliver consistent operations, security, and data governance tools across Hadoop clusters. Hadoop Map Reduce allows various languages to integrate such as Python, C++ etc, . The proposed work is achieved using python a open source scripting language executed in the MapReduce frame work using Hadoop streaming interface, a utility that comes with the Hadoop distribution. This utility allows us to create and run MapReduce jobs with any script as the mapper and the reducer. The mapper reads the data through STDIN a utility in Hadoop streaming and sends the mapped key value pairs to the reducer and the result from the reducer task is stored in the HDFS using STDOUT.

### 4.3.5 Flume

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.

Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS. After the installation of VMWRE and Hadoop for single node next step come the installation of FLUME. For this you need to log in to twitter. After that go to apps on twitter and create a new application. After you agree with all terms and conditions you will got new application. Then set Consumer Key, Consumer Secret, Owner Key and Owner Secret ID. Now access token need to be created. After the creation of access token and refresh you will get all the 4 information. Now you Go to flume home and download Apache Flume.

Flume is composed of the following components.
**Flume Event:** It is the main unit of the data that is transported inside the **Flume** (Typically a single log entry). It contains a payload of the byte array that is to be transported from the source path to the destination path which could be accompanied by optional headers.

**Flume Agent:** Is an independent Java virtual machine daemon process which receives the data (events) from clients and transports to the subsequent destination (sink or agent). **Source:** Is the component of Flume agent which receives data from the data generators say, twitter, Facebook, weblogs from different sites and transfers this data to one or more channels in the form of Flume event. The external source sends data to Flume in a format that is recognized by the target Flume source. Example, an Avro Flume source can be used to receive Avro data from Avro clients or other Flume agents in the flow that send data from an Avro sink, or the Thrift Flume source will receive data from a Thrift sink, or a Flume Thrift RPC client or Thrift Clients are written in any language generated from the Flume thrift protocol. **Channel:** Once, the Flume source receives an Event, it stores this data into one or more channel and buffers them till they are consumed by sinks. It acts as a bridge between the sources and sinks. These channels are implemented to handle any number of sources and sinks.

**Sink:** It stores the data into the centralized stores like HDFS and HBase. **Streaming Twitter Data:** To stream data to our database from twitter we should have the following pre-requisites.

- Twitter account
- Hadoop cluster

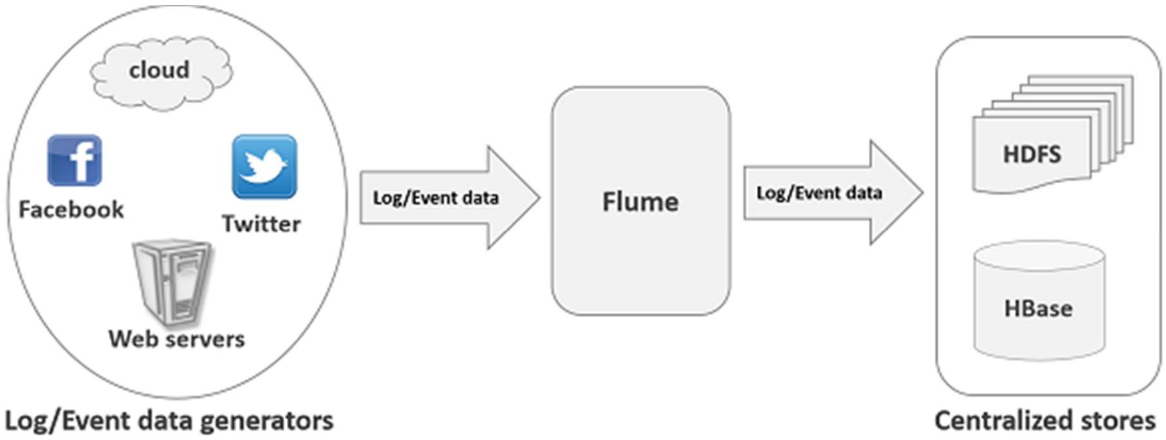


**Fig 4.4: Flume Architecture**

Starting the Agent

Now that we understand the configuration of our source, channel and sink, we need to start up the agent to get the dataflow running. Before we actually start the agent, we need to set the agent to have the appropriate name as defined in the configuration.

The file /etc/default/flume-ng-agent contains one environment variable defined called FLUME_AGENT_NAME. In a production system, for simplicity, the FLUME_AGENT_NAME will typically be set to the hostname of the machine on which the agent is running. However, in this case, we set it to TwitterAgent, and we're ready to start up the process.

### 4.3.6 <u>Hive</u>

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy. Apache Hive (HiveQL) with Hadoop Distributed file System is used for Analysis of data. Hive provides a SQL-like interface to process data stored in HDP. Due its SQL-like interface, Hive is increasingly becoming the technology of choice for using Hadoop.

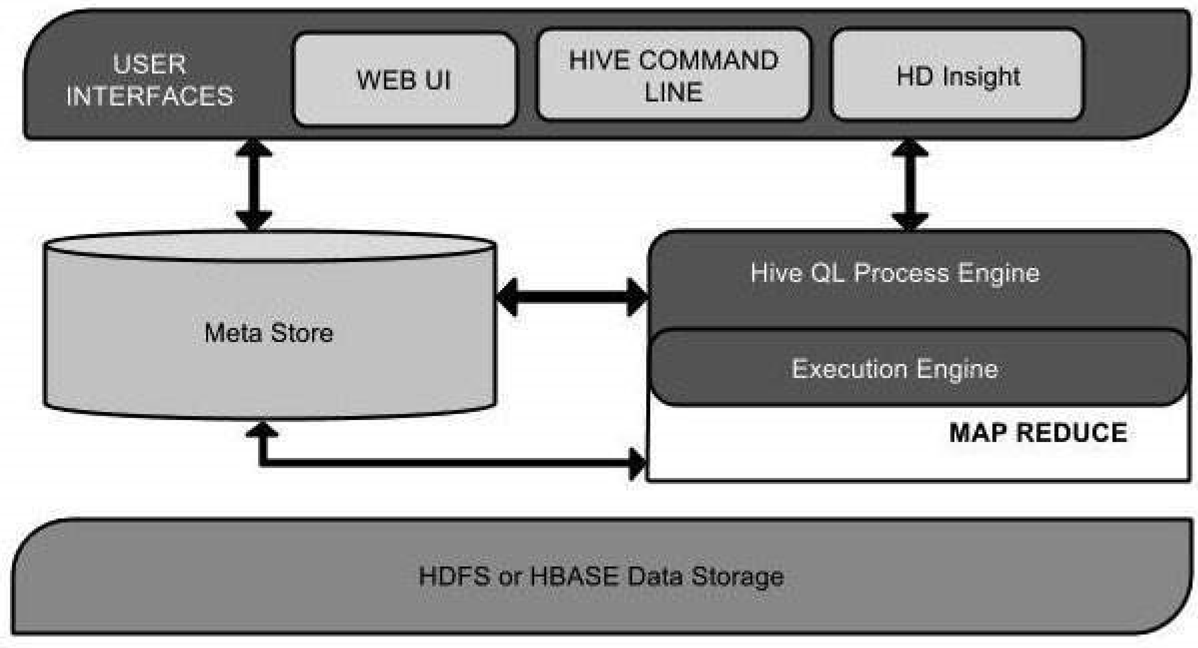


**Fig 4.5: Hive Architecture**

## Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.

## 4.4 System Architecture

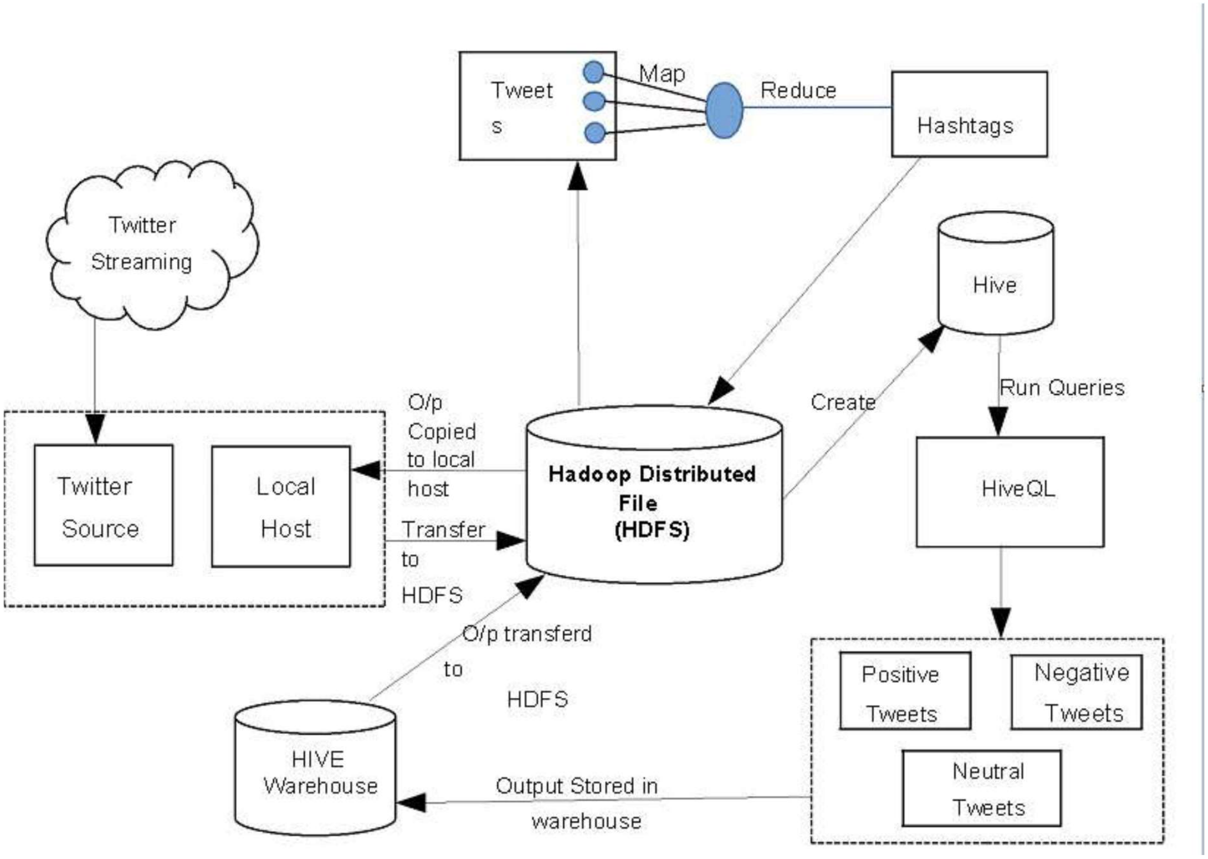


**Fig 4.6: System Architecture**