**Table of contents:**

|  |  |  |
| --- | --- | --- |
| **1** | **Introduction** |  |
| **2** | **Dataset Details** |  |
| **3** | **Related Theory and Facts** |  |
| **4** | **Architecture and Design of System** |  |
| **5** | **Technology Stack** |  |
| **6** | **Implementation Details** |  |
| **7** | **Testing Details** |  |
| **8** | **Limitation of System** |  |
| **9** | **Conclusion and Future Scope** |  |
| **10** | **Bibliography** |  |
| **11** | **Appendix** |  |

Introduction:

Domain:

On November 8th, India’s Prime Minister announced that 86% of the country’s currency would be rendered null and void in 50 days and it will withdraw all 500- and -1000-rupee notes. The country’s most popular currency denominations from circulation, while a new 2,000 rupee note added in. It was posited as a move to crackdown on corruption and virtually untaxed grassroots economy.

Motivation:

The step of **demonetization** by government brought many changes to human lives. Now-a-days social media websites have become the basic medium of expressing the emotions towards daily lives issues and events. The opinions and sentiments of people play an important role in analysing how a sudden change in country can affect a human’s life exhaustively. This motivates us to know about mental stability of individual towards such condition. It has become extremely important to analyse and detect the positive, neutral and negative thoughts of a citizens towards a big decision for a country by a Prime Minister.

Type of solution:

As we know there are many social media platforms to show case our mental perceptions and ideas, from all one of the major network space is Twitter. On Twitter, person’s post is called as tweet and daily millions of tweets which are posted by millions all over the world can be used to analyse consumers’ opinions about a specific situation. In the recent years, these tweets have proven to be a valuable source of information, which are important for the success of any brand, business or the career of a politician also views and thoughts for any situation and decision related to country. In our project, we have adopted Sentiment Analysis of tweets during the demonetization period with an approach to extract positive, neutral and negative tweets by using parts of the speech.

Problem Statement:

To study the sentiments of citizens of our country in the time of “Demonetization” which is sudden and drastic change to a humankind of India.

Dataset Details:

About dataset:

We are using two dataset which have a twitter data, collected at the time demonetization. The datasets are taken form Kaggle.

1. Demonetization\_data29th -

Link to the dataset online: <https://www.kaggle.com/shan4224/demonetization-in-india>

Number of records: 6268

Number of columns: 40

Size of the data: 38.6mb

Important attributes: created\_at\_text, content, from\_user\_location, from\_user\_created\_at, retweet\_count

1. Demonetization\_data23rd –

Link to the dataset online: <https://www.kaggle.com/arathee2/demonetization-in-india-twitter-data>

Number of records: 14940

Number of columns: 16

Size of the data: 5.01mb

Important attributes: created, text, retweetCount, isRetweet

**Related theory and facts:**

Domain Theory:

**Demonetization:** The **demonetization of ₹500 and ₹1000** banknotes was a step taken by the **Government of India** on 8 November 2016, ceasing the usage of all ₹500 and ₹1000 banknotes of the Mahatma Gandhi Series as a form of legal tender in India from 9 November 2016. The announcement was made by the Prime Minister of India **Narendra Modi** in an unscheduled live televised address to the nation at 20:15 Indian Standard Time (IST) the same day. In the announcement, Modi declared circulation of all ₹500 and ₹1000 banknotes of the Mahatma Gandhi Series as invalid and announced the issuance of new ₹500 and ₹2000 banknotes of the Mahatma Gandhi New Series in exchange for the old banknotes.

**Effect of demonetization:** Demonetization leads to cash shortages in the country which proves detrimental to a number of small business, agriculture and transportation. The shortage of cash led to chaos and most people faced problems to exchange their banknotes due to long queues outside banks and ATMs across the country.

About Big Data tool used for project:

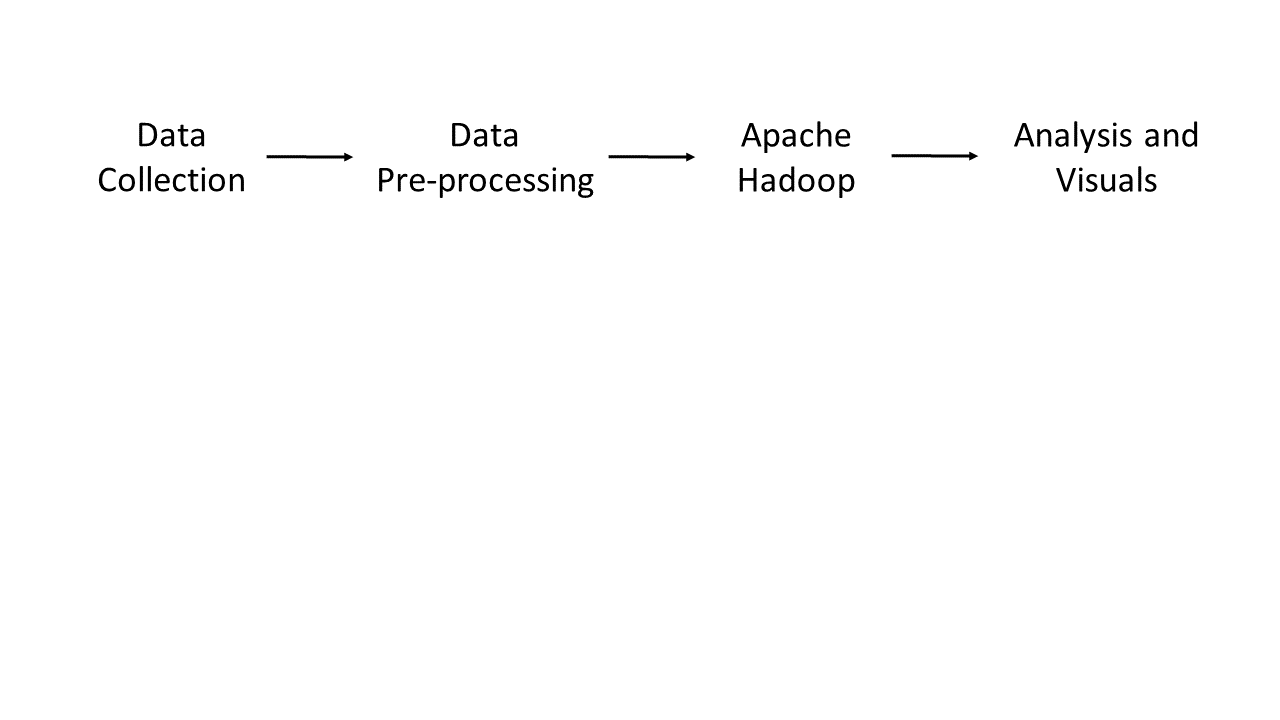
**Hadoop:** Hadoop is one of the best tool options for twitter data analysis as it works for distributed big data, streaming data, time stamped data, text data etc.To analyse and understand the activity occurring on such a massive scale, a relational SQL database is not enough. Such kind of data is well suited to a massively parallel and distributed system like Hadoop.

The tool used here is Pig and Hive. We are not using MapReduce algorithm, since we want to do analysis on complete data. Also, not using aggregated measures as major focus is on text data. Hive and Pig are the components of Hadoop ecosystem. These are high level data flow languages. MapReduce is the inner most layer of Hadoop ecosystem

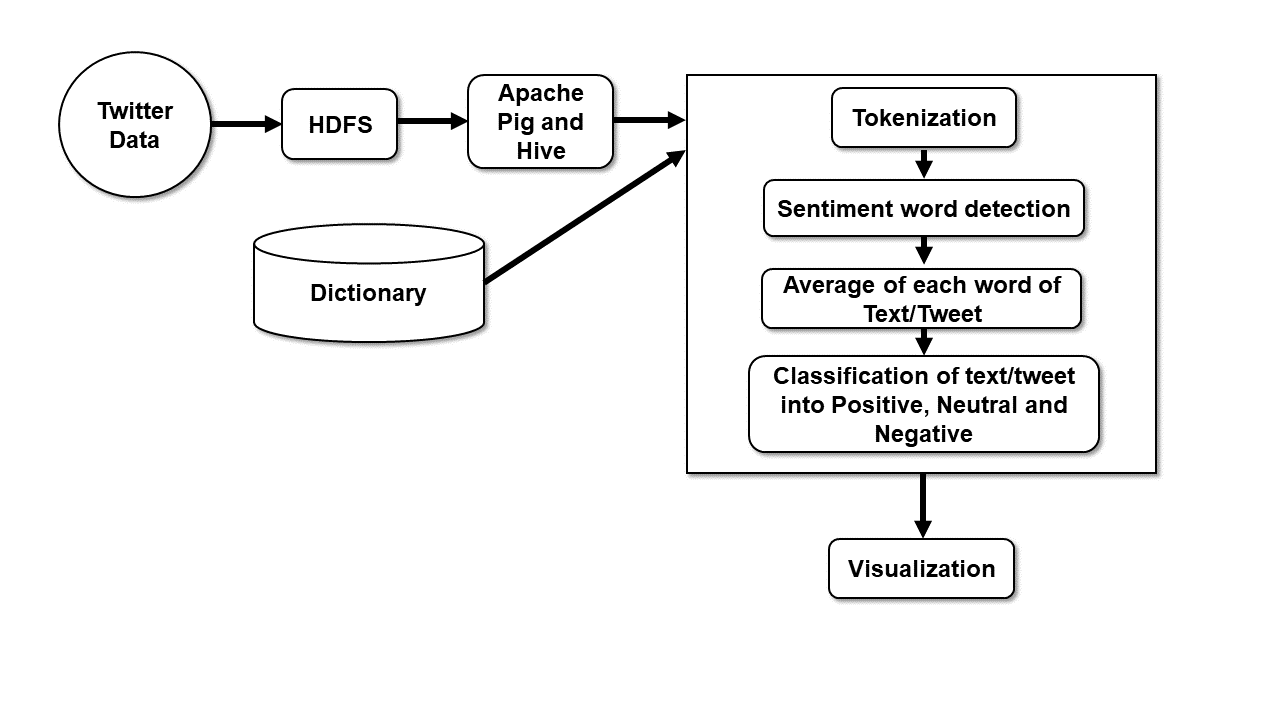
1. **Pig**: As pig is a dataflow language that is built on top of Hadoop to make it easier to process, clean and analyse large amount of data without having to write vanilla MapReduce jobs in Hadoop.
2. **Hive**: Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy. Hive provides the ability to store large amounts of data in HDFS. Hive is an excellent tool for analytical querying of historical data.

**Architecture and Design of the System:**

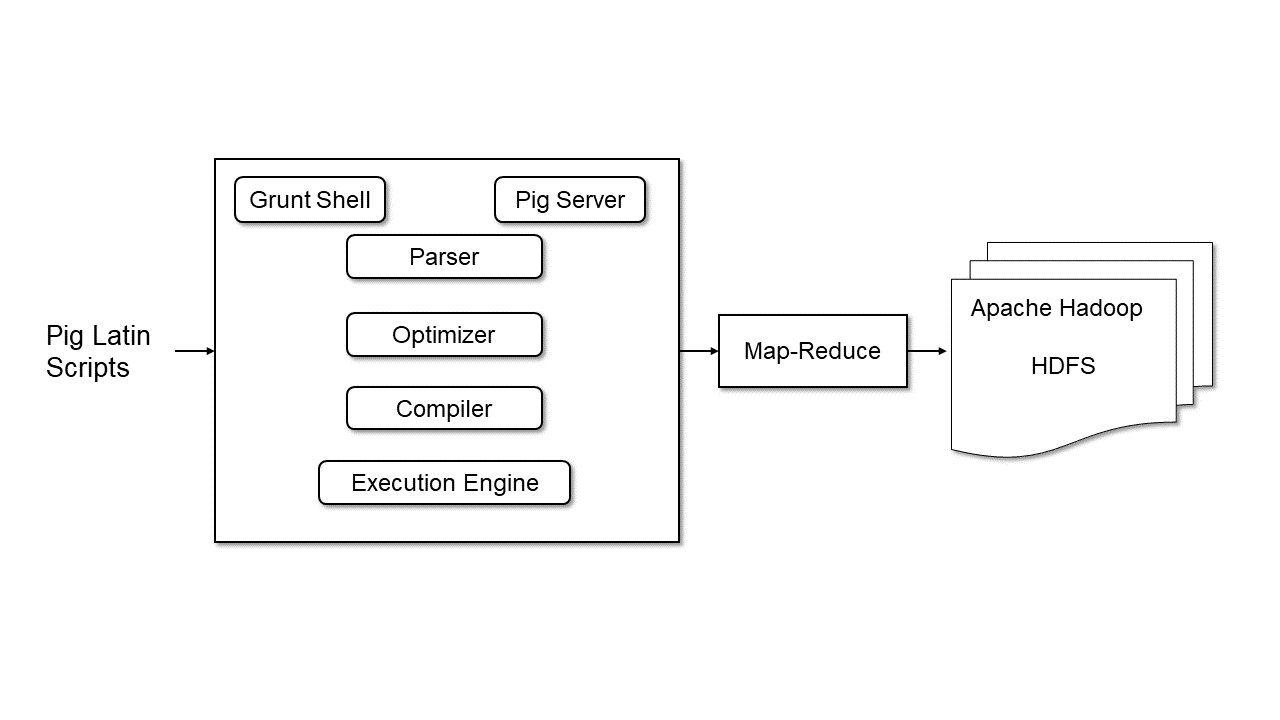
1. Process Flow:



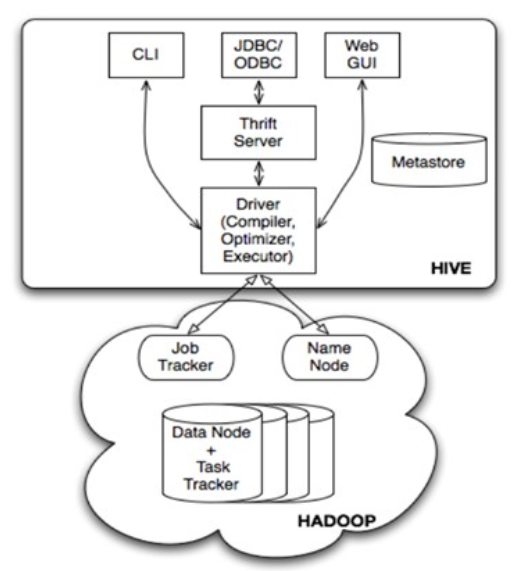
1. Architecture Design:

****

1. Pig framework:



1. Hive framework:



Characteristics:

1. Hive queries are implicitly converted to the mapper and reducer job. Also, extensible, scalable and fast which helps in analysis.
2. Unstructured or semi-structured data can not be problem as Pig is used.
3. Hadoop itself is fault-tolerance.

**Technology Stack:**

1. Big data tool:
   1. Hadoop:
      1. Pig
      2. Hive
2. Framework:
   1. Using main core components of Hadoop framework, MapReduce and HDFS.
      1. MapReduce: analysis is done using the MapReduce framework
      2. HDFS: Hadoop is used for storage and text classification
3. Visualization tool:
   1. Tableau
   2. Python
   3. PowerBI
4. Hardware and Software Specifications used:
   1. Hardware:
      1. Personal Computer/Laptop
      2. Router for internet connection
   2. Software:
      1. VM-ware
      2. Cloudera virtual machine
      3. Anaconda for python: Libraries like pandas, matplotlib, seaborn
      4. Tableau and PowerBI
      5. MS-Excel

**Dependencies:**

1. A system with good RAM and processor for Hadoop.
2. Internet.
3. Large dictionary for optimal results.

**Limitations:** The AFINN Dictionary have only 2500 words which may not be enough to get complete analysis report and may develop to wrong result.

**Conclusion:** Taking up the “Demonetization” time for our sentimental analyse of human emotions as it was a time people were affected the most emotionally and economically. Using people’s social posted text we tried to know how much this situation have positive or negative effects. We were able to find out the sentiments of the people based on the tweets they posted on twitter.

**Future work:** To create a model like textblob and NLTK. Will make it possible to know more about effects of “Demonetization” on economy.