# **Big Data Programming**

**Project Increment 1 Report**

**Title: Analysis on Covid-19 Vaccination**

**Team Members:**

Prakhya Srungarapu - [psrv7@mail.umkc.edu](mailto:psrv7@mail.umkc.edu)

Sai Saranya Vipparla - [svv7x@mail.umkc.edu](mailto:svv7x@mail.umkc.edu)

Vamshi Krishna Routhu - [nrqnh@mail.umkc.edu](mailto:nrqnh@mail.umkc.edu)

Shiva Kumar Murarishetti – [smpf4@mail.umkc.edu](mailto:smpf4@mail.umkc.edu)

**Introduction:**

Over the next few months, COVID-19 Vaccine could become available for use all over the world. The distribution of vaccinations to states and localities, as well as monitoring their distribution to individuals can easily be done by having the proper analysis of people data who are actually in need of vaccination. Ensure that enough vaccines are administered would be a historically challenging task. However, million doses of vaccination may be required to distribute among people depending on the period and intensity of immunization, there are still many concerns from people regarding the vaccination reaching out to public which will be clearly implemented in our project in form of data analysis.

**Motivation:**

As large number of people are in need of vaccination, we need to have proper data insight so that vaccines can be distributed in an effective and fastest way.

**Objectives:**

To analyze number of people across the world who got vaccinated, people who are interested to get vaccinated, Side effects of people who got vaccinated.

To analyze the data based on several factors such as age, gender, region, health conditions. We will be using different Hadoop ecosystem tools to collect the data, store, analyze and generating visualizations.

**Features:**

The features of this project would be analysis on the twitter data, public data sets available. Storing the data in the HDFS, Processing data using Spark, Querying using Hive Query language. Generating visualizations using matplotlib, NumPy, Pandas, Seaborn and Tableau. Showing the visualizations on the web application. Deploying the application in the docker.

**Significance:**

* As many countries are unable to produce the vaccine for their population, their government should have the track of the people getting vaccinated, people showing the interest towards vaccination.
* How does their country economy get impacted due to vaccination?
* How the distribution should be done?
* Big data tools help to process huge amounts of data efficiently.
* This would be helpful for the governments and the common people to make their decision towards the vaccination.

**Datasets:**

1. Twitter streaming Data - Using Twitter developer API

Create twitter developer account, generate API key, and token which can be used to download tweets using Tweepy.

1. Static datasets currently available on internet

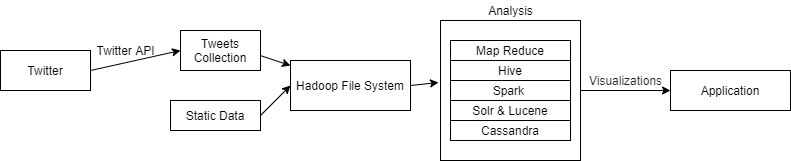
<https://covidtracking.com/about-data/data-summary>

<https://www.kaggle.com/c/stanford-covid-vaccine>

**Architecture:**



**Workflow:**



**Analysis:**

**Project Management:**

Implementation Status Report:

Work Completed:

* Project Outcome - Prakhya Srungarapu
* Investigating the Datasets - Sai Saranya Vipparla
* Planning the queries to be implemented - Vamshi Krishna Routhu
* Design of Project - Shiva Kumar Murarishetti

Work to be Completed:

* Spark SQL Queries, Data Frames creation
* Sentimental Analysis
* Hive QL Queries
* Generating visualizations
* Building Web Application using Django Framework
* Deploying in Docker

**Story Telling:**

**Who?**

Since the beginning of vaccine delivery in the United States on December 14, more than 68 million doses have been distributed, totaling 13.9 million. Government and public health departments will use this to monitor vaccine progress using data and insights, helping them to increase the efficacy of their programs. This will help healthcare providers streamline their vaccine processes.

**What?**

Having improper data of the COVID vaccination would hinder the process of improving the vaccine's efficiency, knowing the total vaccinations given out with their trail-based success rate would help to further improve it.

**When?**

COVID-19 vaccination will be an important tool to help stop the pandemic. In real time, knowing how many people have taken the vaccine around the globe helps us to get an analysis of how well the vaccine is being received by the people and to know the distribution structure of the vaccine.

**Where?**

As Covid is a global problem this would be effective and helpful for most sectors but mostly would be beneficial for the health and research departments which are looking to develop immunization for various new diseases and infections.

**Why?**

Following the isolation of a coronavirus in December 2019, its genetic sequence was released on January 11, 2020, prompting an immediate international response to plan for an outbreak and accelerate the production of a COVID-19 vaccine. Right now, orchestrating the delivery of billions of vaccine doses is the greatest problem the world faces. In ensuring that it is achieved safely, efficiently, and equitably, technology will play a critical role. Tracking the vaccination progress will help in analyzing the number of people vaccinated and how many new vaccines needs to be developed.

**How?**

People understandably have a lot of questions about the new vaccines, particularly because vaccines are being developed at an alarming pace and the virus is still revealing its secrets. Health experts stress the importance of vaccinating as many individuals as possible to avoid the outbreak, as global COVID-19 cases exceed 100 million. The more circulating virus indicates the greater risk of mutations, and some recent strains tend to be more infectious already, making vaccination even more of a race against time**.**

**References:**

1. <https://www.kaggle.com/c/stanford-covid-vaccine>
2. <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/index.html>
3. <https://hadoop.apache.org/>
4. <https://hive.apache.org/>
5. <https://spark.apache.org/>
6. <https://cassandra.apache.org/>
7. <https://lucene.apache.org/solr/>
8. <https://developer.twitter.com/en>
9. <https://www.djangoproject.com/>
10. <https://www.docker.com/>
11. <https://www.cloudera.com/products/open-source/apache-hadoop.html>
12. <https://covidtracking.com/analysis-updates/how-we-hope-vaccines-will-be-tracked>