# **Covid – 19: Vaccines Tweets Globally (Real-Time Processing)**

What is Real-Time Processing?

* Real-Time/Streaming data Processing where the “Data Flow” is continuous & which is processed immediately after the retrieval with the sole aim to extract insights and useful trends out of it.

For e.g., *IoT data.*

**Context:**

Twitter is an online Social Media Platform where people share their thoughts as tweets. It is observed that some people misuse it to tweet hateful content. Twitter is trying to tackle this problem.

**Dataset:**

**Tech-Stack:**

1. Python\*
2. Anaconda(IDE) \*
3. Spark/PySpark\*
4. Apache (Flume\*, Flink, STORM), & Spark Streaming\*
5. Hadoop(HDFS & YARN\*)
6. Hive Serde\*

**High-Level Diagram:**

**Objective:**

* Milestone – 1
* Setup Twitter developer by creating an account.
* Use Twitter API V2.0
* Milestone – 2
* Setup Apache Flume locally.
* Connect Flume with a Twitter account.
* **Flume no longer works with most tutorials because the twitter ingestion that flume uses (even the latest version) is using the old v1.1 streaming API, which is now deprecated, even with Elevated Access Deprecation announcement** ([Deprecation announcement: Removing compliance messages from statuses/filter and retiring statuses/sample from the Twitter API v1.1 - Announcements - Twitter Developers (twittercommunity.com)](https://twittercommunity.com/t/deprecation-announcement-removing-compliance-messages-from-statuses-filter-and-retiring-statuses-sample-from-the-twitter-api-v1-1/170500/4) )
* *Alternate method we tried to use to collect the “Tweets” from twitter using tweepy, unfortunately pulling of tweets has been blocked from Tweeter itself.*
* Milestone – 2.1
* Download the Tweet dataset from Kaggle.
* Dump the dataset into HDFS.
* Milestone – 3
* Setup Apache hive in your local
* 1. Java Version - 8
* 2. Hadoop – 2.7.0
* 3. Apache Derby

Create an external table on an HDFS file (You can use Hive Serde because HDFS data would of json type).

Write a query to display, how many tweets have been made regarding the keyword ‘Covid-19 Vaccines in each country.

**Ref Link:**

[Batch Processing vs Real Time Processing - Comparison - DataFlair (data-flair.training)](https://data-flair.training/blogs/batch-processing-vs-real-time-processing/)

[What is Big Data Streaming? - Definition from Techopedia](https://www.techopedia.com/definition/31752/big-data-streaming)

[Fetching Twitter Data into HDFS (Hadoop) using Apache Flume .....Step By Step Guide on Windows 10 - YouTube](https://www.youtube.com/watch?v=7yKaQDccL7Y)

[Apache Flume - Fetching Twitter Data (tutorialspoint.com)](https://www.tutorialspoint.com/apache_flume/fetching_twitter_data.htm)

[COVID-19 All Vaccines Tweets | Kaggle](https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets)