# **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.

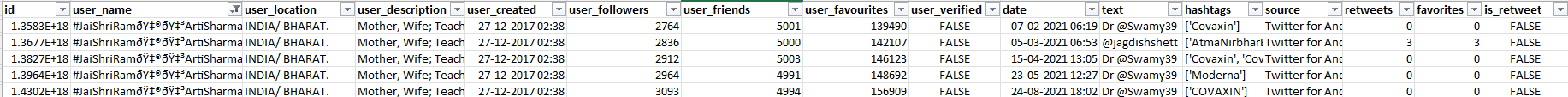
For e.g., *IoT data.*

**Context:**

Twitter is an online Social Media Platform where people share their thoughts as tweets.

**Dataset:**

* Datasets consist of **16-columns** and **2,28,208** records.
* Datasets was created in the year **Dec 2020 – Nov 2021**

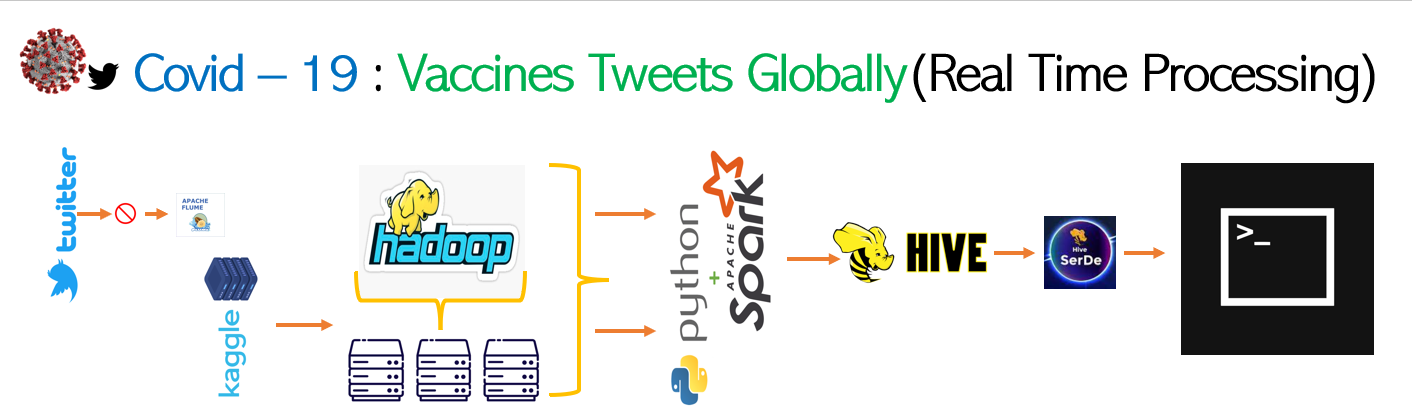


**Tech-Stack:**

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

\* Technology Used

**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

* + Milestone – 3.1
    - Do the minimal transformation.

1. Drop
   * Milestone – 3.2
     + Create a “Temp” directory in HDFS.
     + Use Python to covert the “CSV” to “JSON/Stream Data” format in a “single file”.
     + Push the data in the “Temp” directory using PySpark.

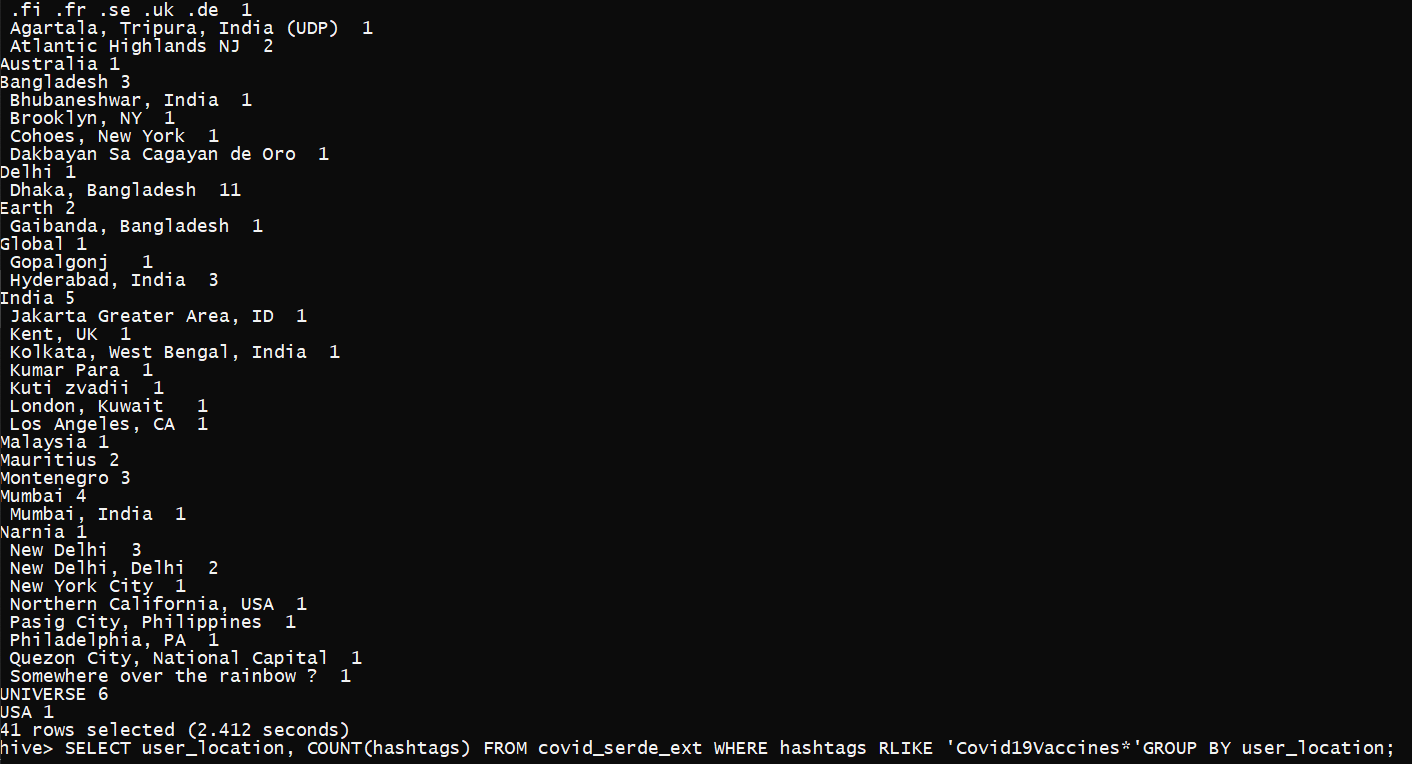
* Milestone – 4
* Create an external table on an HDFS file (use Hive Serde because HDFS data would of JSON type).
* Milestone – 4.1
* Create an “Internal Table” with schema in Hive.
* Load the data from HDFS Dir to the Hive Table.
* Milestone – 4.2
* Create “External Table” using Hive SerDe.
* Append the data from Hive Table to external table using Hive SerDe as the data is in JSON Format.
* Milestone – 4.3
* How to create an external table using Hive SerDe:

A black rectangular object with white border

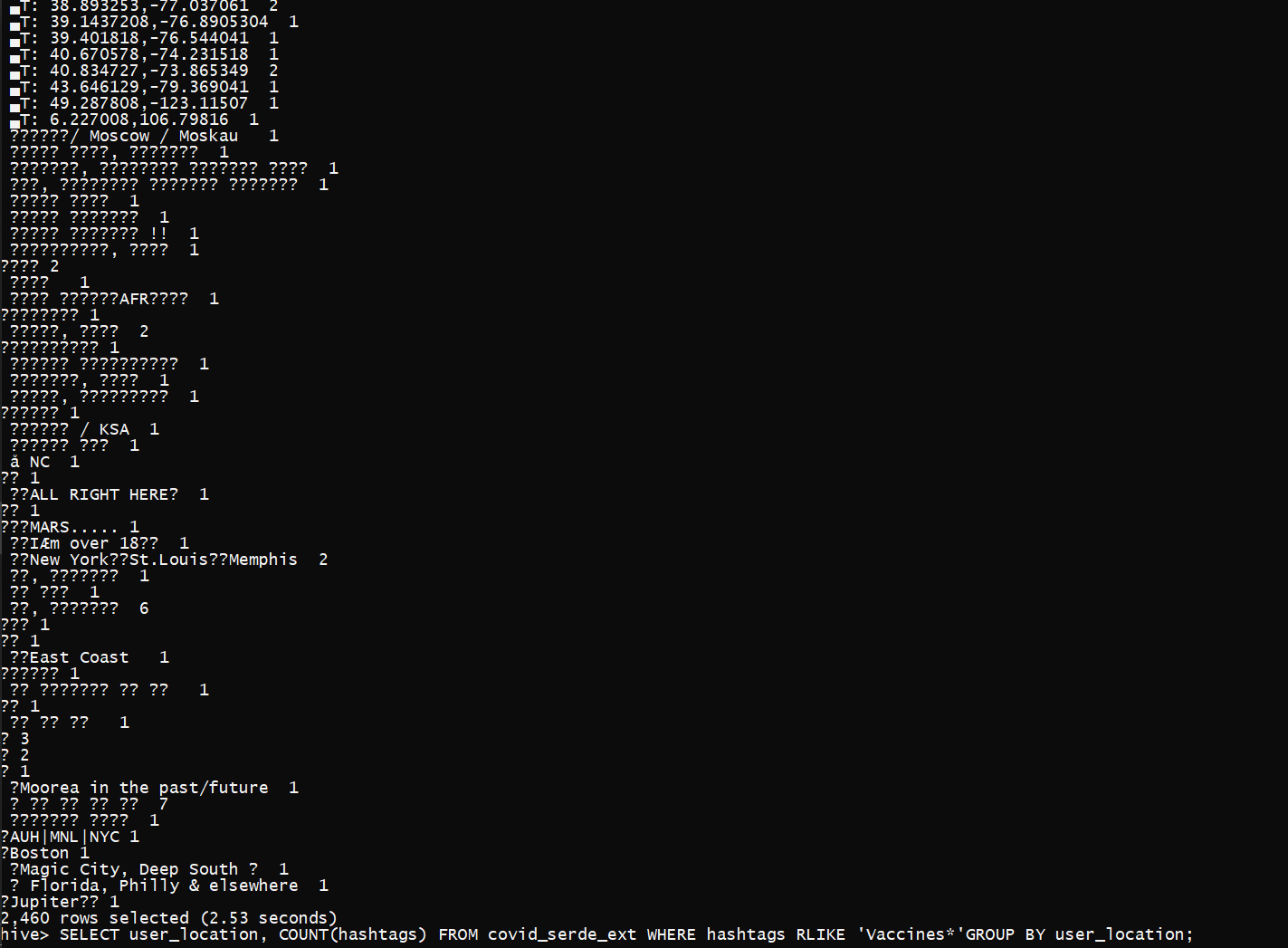
Description automatically generated

* Loading the data from HDFS into Hive Table:
* LOAD DATA INPATH ‘Path’ INTO TABLE “Table\_name”;
* Milestone – 5
* Write a query:

1. How many tweets have been made regarding the keyword ‘Covid-19 Vaccines’ in each country?



1. How many tweets have been made regarding the keyword ‘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)

[apache spark - Pyspark dataframe write to single json file with specific name - Stack Overflow](https://stackoverflow.com/questions/43269244/pyspark-dataframe-write-to-single-json-file-with-specific-name)

[python - How to change csv file name while writing in spark? - Stack Overflow](https://stackoverflow.com/questions/68810887/how-to-change-csv-file-name-while-writing-in-spark)

[scala - Write single CSV file using spark-csv - Stack Overflow](https://stackoverflow.com/questions/31674530/write-single-csv-file-using-spark-csv)

[OSError: Prior attempt to load libhdfs failed - Google Search](https://www.google.com/search?q=OSError%3A+Prior+attempt+to+load+libhdfs+failed&sxsrf=AB5stBjB1n4VU7gtIWjRbNxEHV2jBcbQ4w%3A1688880534961&source=hp&ei=lkWqZPTDN6zcseMPveiaeA&iflsig=AD69kcEAAAAAZKpTppxKMdVHCT__6CrsTVr7kdN4z6Bi&ved=0ahUKEwj05uei8oCAAxUsbmwGHT20Bg8Q4dUDCAk&uact=5&oq=OSError%3A+Prior+attempt+to+load+libhdfs+failed&gs_lcp=Cgdnd3Mtd2l6EAMyBwgjEK4CECdQAFgAYK0FaABwAHgAgAHrAYgB6wGSAQMyLTGYAQCgAQKgAQE&sclient=gws-wiz)

[hadoop - Hive ParseException - cannot recognize input near 'end' 'string' - Stack Overflow](https://stackoverflow.com/questions/18640760/hive-parseexception-cannot-recognize-input-near-end-string)

[Hive SERDE (dbmstutorials.com)](https://dbmstutorials.com/hive/hive-serde.html)

[Hadoop Lessons: Loading data into Hive Table](http://www.hadooplessons.info/2014/12/loading-data-into-hive-table.html)