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

What is Real-Time Processing?

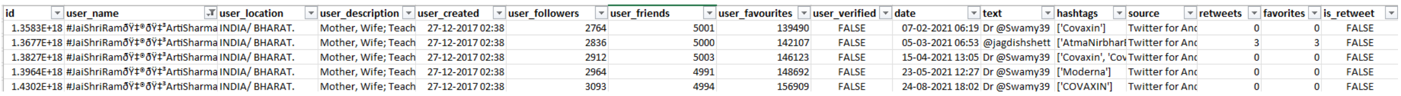
- Real-Time/Streaming data Processing where the "Data Flow" is continuous & which is processed immediately after the retrieval with the sole aim to extract insights and useful trends out of it.
  For e.g., *IoT data*.

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**



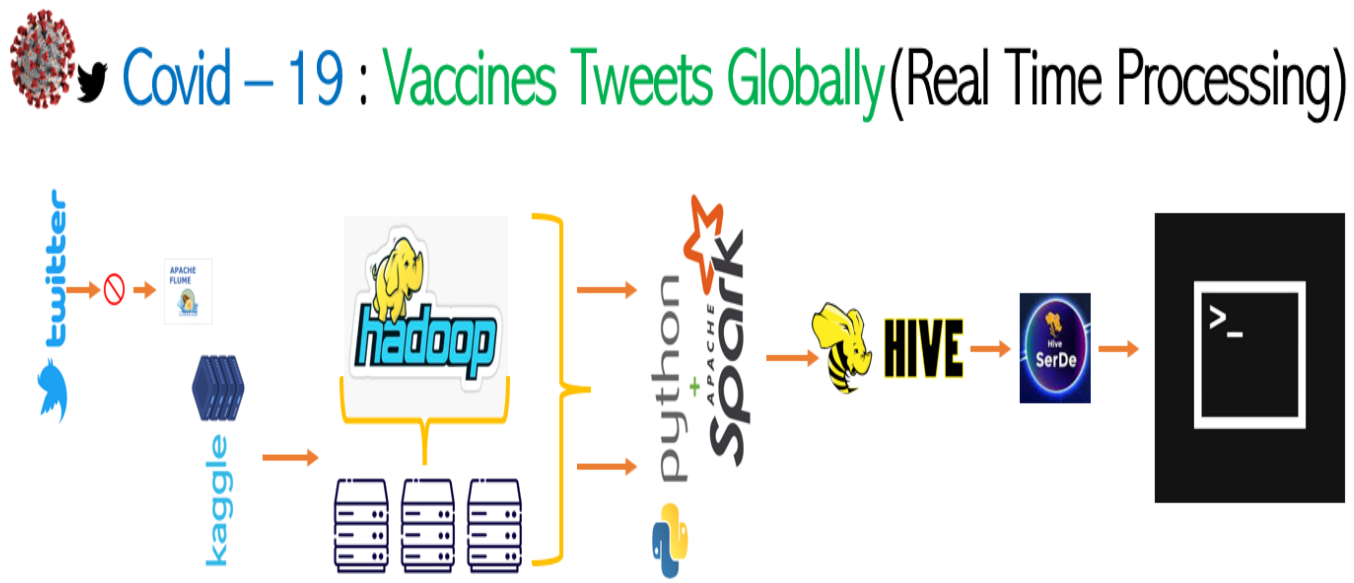
id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	favorites	is_retweet
1.3583E+18	#JaiShriRamðŸ‡®ðŸ‡³ArtiSharma	INDIA/ BHARAT.	Mother, Wife; Teach	27-12-2017 02:38	2764	5001	139490	FALSE	07-02-2021 06:19	Dr @Swamy39	['Covaxin']	Twitter for Anc	0	0	FALSE
1.3677E+18	#JaiShriRamðŸ‡®ðŸ‡³ArtiSharma	INDIA/ BHARAT.	Mother, Wife; Teach	27-12-2017 02:38	2836	5000	142107	FALSE	05-03-2021 06:53	@jagdishshett	['AtmaNirbhar	Twitter for Anc	3	3	FALSE
1.3827E+18	#JaiShriRamðŸ‡®ðŸ‡³ArtiSharma	INDIA/ BHARAT.	Mother, Wife; Teach	27-12-2017 02:38	2912	5003	146123	FALSE	15-04-2021 13:05	Dr @Swamy39	['Covaxin', 'Cov	Twitter for Anc	0	0	FALSE
1.3964E+18	#JaiShriRamðŸ‡®ðŸ‡³ArtiSharma	INDIA/ BHARAT.	Mother, Wife; Teach	27-12-2017 02:38	2964	4991	148692	FALSE	23-05-2021 12:27	Dr @Swamy39	['Moderna']	Twitter for Anc	0	0	FALSE
1.4302E+18	#JaiShriRamðŸ‡®ðŸ‡³ArtiSharma	INDIA/ BHARAT.	Mother, Wife; Teach	27-12-2017 02:38	3093	4994	156909	FALSE	24-08-2021 18:02	Dr @Swamy39	['COVAXIN']	Twitter for Anc	0	0	FALSE

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

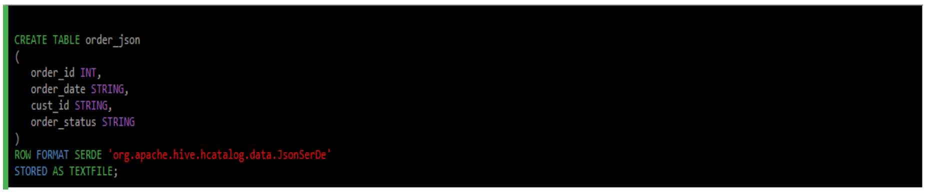
High-Level Diagram:



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))
    - *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:



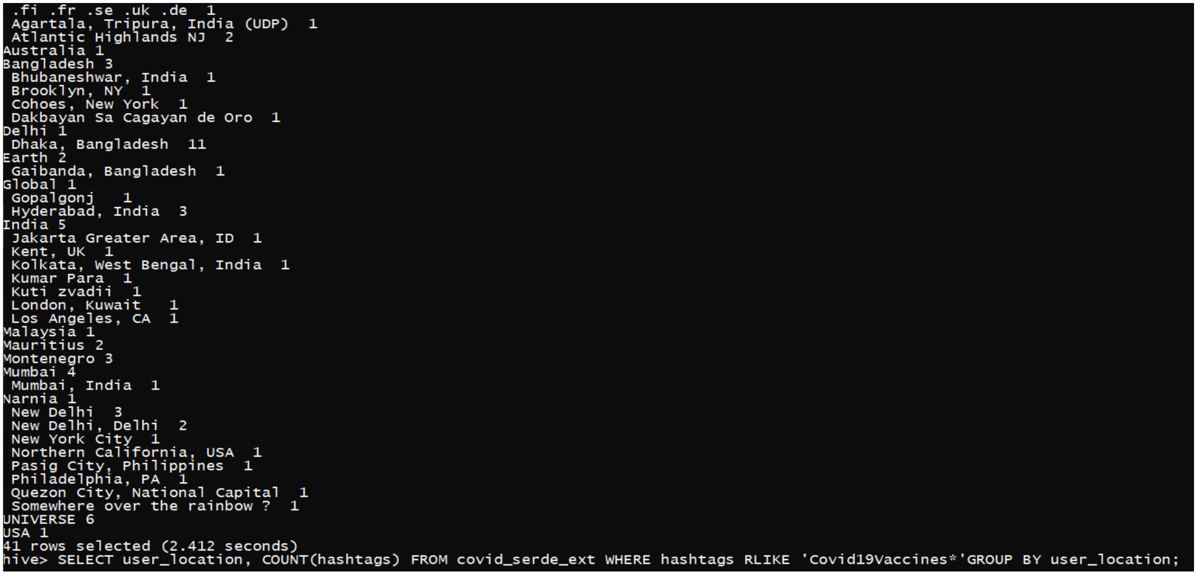
```
CREATE TABLE order_json
(
   order_id INT,
   order_date STRING,
   cust_id STRING,
   order_status STRING
)
ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
STORED AS TEXTFILE;
```

- Loading the data from HDFS into Hive Table:
- LOAD DATA INPATH 'Path' INTO TABLE "Table_name";

Milestone – 5

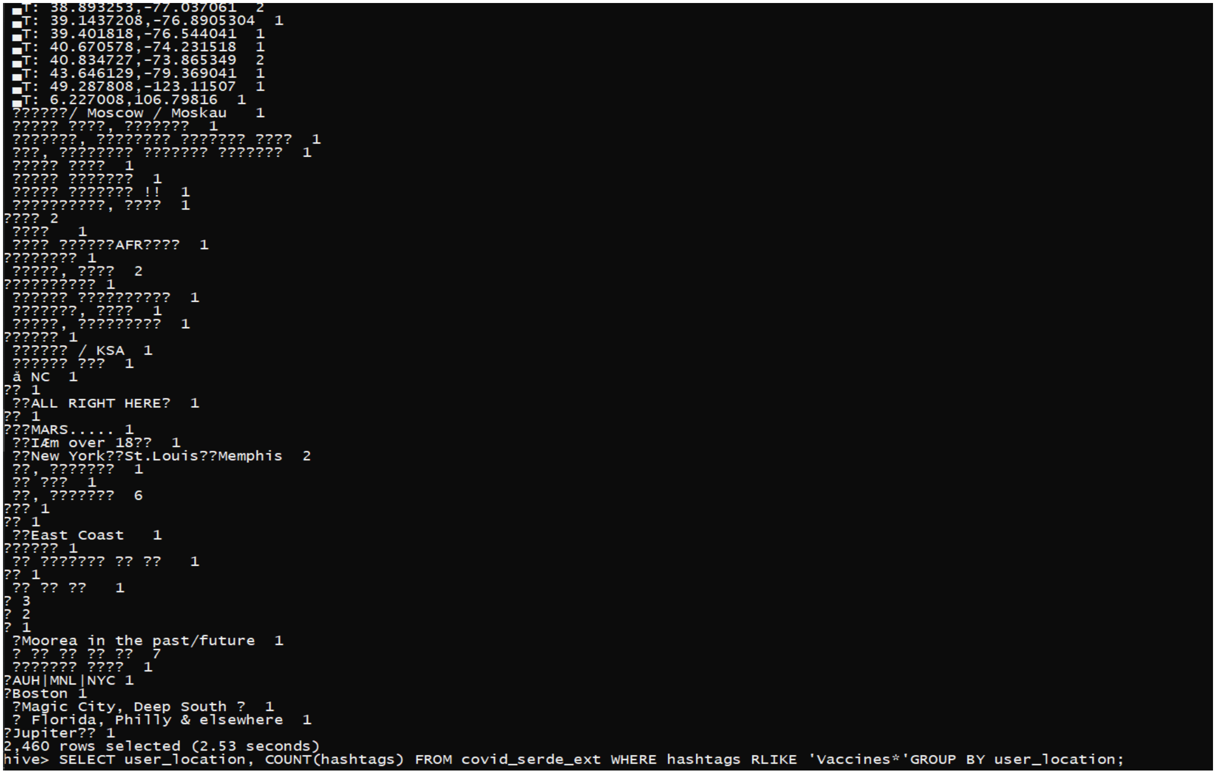
- Write a query:

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



```
 .fi .fr .se .uk .de  1
 Agartala, Tripura, India (UDP)  1
 Atlantic Highlands NJ  2
Australia 1
Bangladesh 3
 Bhubaneshwar, India  1
 Brooklyn, NY  1
 Cohoes, New York  1
 Dakbayan Sa Cagayan de Oro  1
Delhi 1
 Dhaka, Bangladesh  11
Earth 2
 Gaibanda, Bangladesh  1
Global 1
 Gopalgonj   1
 Hyderabad, India  3
India 5
 Jakarta Greater Area, ID  1
 Kent, UK  1
 Kolkata, West Bengal, India  1
 Kumar Para  1
 Kuti zvadii  1
 London, Kuwait   1
 Los Angeles, CA  1
Malaysia 1
Mauritius 2
Montenegro 3
Mumbai 4
 Mumbai, India  1
Narnia 1
 New Delhi  3
 New Delhi, Delhi  2
 New York City  1
 Northern California, USA  1
 Pasig City, Philippines  1
 Philadelphia, PA  1
 Quezon City, National Capital  1
 Somewhere over the rainbow ?  1
UNIVERSE 6
USA 1
41 rows selected (2.412 seconds)
hive> SELECT user_location, COUNT(hashtags) FROM covid_serde_ext WHERE hashtags RLIKE 'Covid19Vaccines*'GROUP BY user_location;
```

2. How many tweets have been made regarding the keyword 'Vaccines' in each country?



```
 ▄T: 38.893253,-77.037061  2
 ▄T: 39.1437208,-76.8905304  1
 ▄T: 39.401818,-76.544041  1
 ▄T: 40.670578,-74.231518  1
 ▄T: 40.834727,-73.865349  2
 ▄T: 43.646129,-79.369041  1
 ▄T: 49.287808,-123.11507  1
 ▄T: 6.227008,106.79816  1
 ??????/ Moscow / Moskau   1
 ????? ????, ???????  1
 ???????, ???????? ??????? ????  1
 ???, ???????? ??????? ???????  1
 ????? ????  1
 ????? ???????  1
 ????? ??????? !!  1
 ??????????, ????  1
???? 2
 ????   1
 ???? ??????AFR????  1
???????? 1
 ?????, ????  2
?????????? 1
 ?????? ??????????  1
 ???????, ????  1
 ?????, ?????????  1
?????? 1
 ?????? / KSA  1
 ?????? ???  1
 ă NC  1
?? 1
 ??ALL RIGHT HERE?  1
?? 1
???MARS..... 1
 ??IÆm over 18??  1
 ??New York??St.Louis??Memphis  2
 ??, ???????  1
 ?? ???  1
 ??, ???????  6
??? 1
?? 1
 ??East Coast   1
?????? 1
 ?? ??????? ?? ??   1
?? 1
 ?? ?? ??   1
? 3
? 2
? 1
 ?Moorea in the past/future  1
 ? ?? ?? ?? ??  7
 ??????? ????  1
?AUH|MNL|NYC 1
?Boston 1
 ?Magic City, Deep South ?  1
 ? Florida, Philly & elsewhere  1
?Jupiter?? 1
2,460 rows selected (2.53 seconds)
hive> SELECT user_location, COUNT(hashtags) FROM covid_serde_ext WHERE hashtags RLIKE 'Vaccines*'GROUP BY user_location;
```

GitHub: https://github.com/dom007-rock/covid_19_streaminng/tree/5b3a81d02bcb9409b8284cc41a70ee20e5a27d4e

Ref Link:

Batch Processing vs Real Time Processing - Comparison - DataFlair (data-flair.training)
What is Big Data Streaming? - Definition from Techopedia
Fetching Twitter Data into HDFS (Hadoop) using Apache Flume .....Step By Step Guide on Windows 10 - YouTube
Apache Flume - Fetching Twitter Data (tutorialspoint.com)
COVID-19 All Vaccines Tweets | Kaggle
apache spark - Pyspark dataframe write to single json file with specific name - Stack Overflow
python - How to change csv file name while writing in spark? - Stack Overflow
scala - Write single CSV file using spark-csv - Stack Overflow
OSError: Prior attempt to load libhdfs failed - Google Search
hadoop - Hive ParseException - cannot recognize input near 'end' 'string' - Stack Overflow
Hive SERDE (dbmstutorials.com)
Hadoop Lessons: Loading data into Hive Table