Covid-19 Cases on World map (Batch processing)

What is Batch Processing?

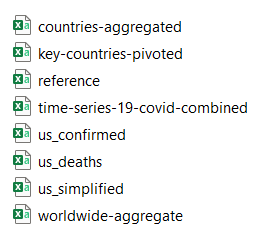
Data which is collected over time and then fed into a system where “meaning” is extracted and further it is stored(warehouses) or it is used to process.

For e.g., *Daily processing of bills in a retail shop*.

**Context:**

* A new coronavirus designated 2019-nCoV was first identified in Wuhan, the capital of China's Hubei province.
* People developed pneumonia without a clear cause and for which existing vaccines or treatments were not effective.
* The virus has shown evidence of human-to-human transmission.
* Transmission rate (rate of infection) appeared to escalate in mid-January 2020
* As of 15th May 2023, approximately 69,27,378 deaths have been confirmed.

**Content:**



**Tech Stack:**

1. Hadoop two main components(HDFS & YARN)
2. Python
3. Anaconda(IDE)
4. Spark/PySpark
5. SQL(PostgreSQL/MySQL)\*
6. NoSQL(Cassandra/HBase)
7. Dashboarding(Tableau, Power-BI, Grafana\*, Kibana)

**High-Level Diagram:**

**Milestones:**

* **Milestone – 1**

- Clone/Download file from Git repo & save it in your local.

- [GitHub - datasets/covid-19: Novel Coronavirus 2019 time series data on cases](https://github.com/datasets/covid-19)

* **Milestone – 2**

- Setup Hadoop in local.

- Here we need to focus on only two main components of Hadoop(HDFS &

YARN)

- HDFS is distributed file system so now copy that file from local to HDFS.

*- hdfs dfs mkdir /dir\_name*

*- hdfs dfs -ls /*

*- hdfs dfs -copyFromLocal “path” /dir\_name*

*- hdfs dfs -ls /dir\_name*

* **Milestone – 3**

- Setup spark in your local.

* We need to install Spark in our local and further we need to instantiate.
* To initiate we need we can use “spark” in the CMD, or we can use it in Jupyter Notebook.

- Write a spark application which reads file from HDFS and creates a DataFrame.

a. With the help of Spark we could pull the data from HDFS(Hadoop Distributed File

System) and create a DataFrame.

* Just looking a quick view, we can understand how many columns are there and we can plan what type of further transformation we could do.
* **Milestone – 4**
* Following basic transformation on the data:

1. Ingesting only the required columns.
2. Dropping duplicate values
3. Null handling transformation for all the columns.

String: NA

INT: 0

Float: 0.0

Date/Timestamp: 1800-01-01

1. Column renaming
2. Type casting according to the required data type.

* **Milestone – 5**
* Aggregate data in spark application.

1. *How many total cases so far?*
2. *How many recovered cases so far?*
3. *How many deaths so far?*

* **Milestone – 5**
* Write the aggregated data in output table using transactional database like MySQL/Postgres or in NoSQL database.

**Ref Link:**

1. [2 Beginner Level Projects For Aspiring Data Engineers ⌛️Covid-19 Data | Real Time Data Processing 👈 - YouTube](https://www.youtube.com/watch?v=GmR2migiu6A)