Data modeling and pipelining

This section tests your data engineering capabilities. You can attempt this section or the next (system design) section of the quiz.

Assume you are developing an application to bring out the latest in COVID related clinical trials. While there are any number of web sites providing dashboards on COVID related trials, none of them provide an easy way for users to get answers to questions of common interest such as:

1. What are the leading candidates for vaccination/treatment as evidenced by:
1. results obtained so far,
2. number and nature of subjects enrolled in the trials,
3. source of funding for the trials,
4. experience of the biotech/pharma company that is the innovator,
5. mechanism of action for the proposed drug/vaccine
2. What are the limitations of leading vaccine/therapy candidates such as:
1. Demography they've been tested on: E.g., age group, gender, race/nationality, social strata
2. Ease of distribution and administration: E.g., number of visits to the clinic needed per patient, storage conditions needed to keep the drug potent and stable, devices needed for administering the medicine etc.
3. Manufacturing capacity available to ramp-up production
3. What are the co-morbidities that complicate COVID treatment and what treatment options are being tested in each case?

In the questions below, you'll be asked to design a data pipeline and a data model needed to create an application that will answer the above questions.

Through this submission, you are expected to demonstrate your understanding of relational data modeling principles and your ability to grasp specifics of the domain you are modeling.

To understand the basics of data modeling, we suggest using a formal text book (such as https://opentextbc.ca/dbdesign01/) or course (such as https://www.coursera.org/learn/database-management/).

To understand the domain to model: Read the descriptions at clinicaltrials.gov for a few trials such as: https://clinicaltrials.gov/ct2/show/NCT04794946. Use the glossary at https://clinicaltrials.gov/ct2/about-studies/glossary for any clinical trials related terms you do not understand.

Question 2:

Describe a data pipeline that will collect the required data from public sources such as but not limited to clinicaltrials.gov, cleanses it and loads the data model you created in answer to the previous question. Illustrate the pipeline using a diagram and describe each component's responsibility, challenges involved and technology choices. Provide here a link to your work.

Data Pipeline:

A data pipeline architecture is an arrangement of objects that extracts, regulates, and routes data to the relevant system for obtaining valuable insights. It embraces the ETL pipeline as a subset.

Three main factors should be considered when building a data pipeline:

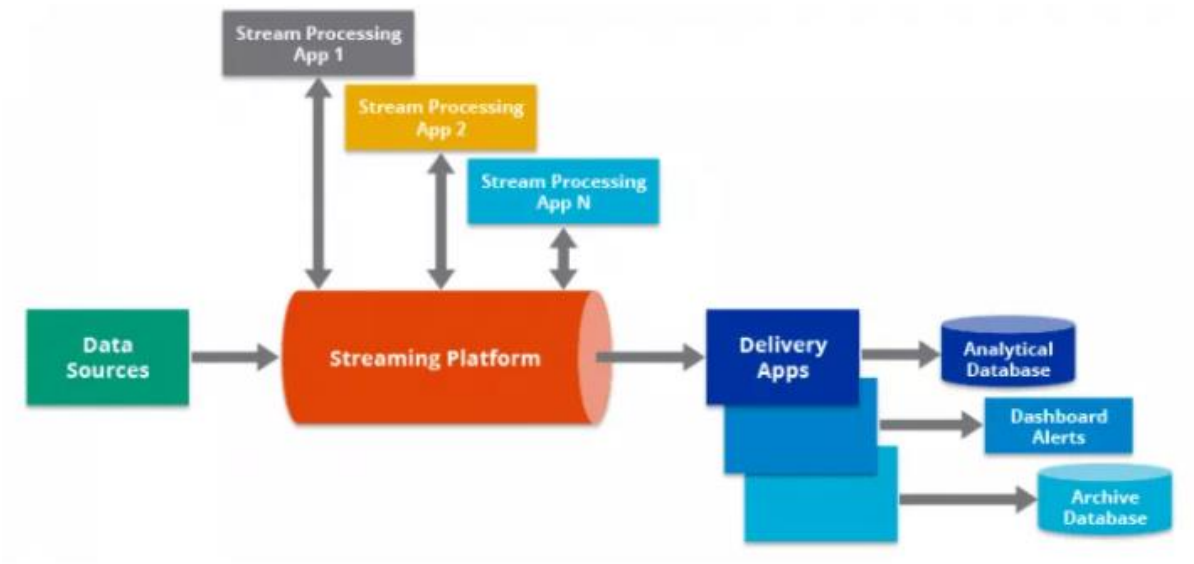
Throughput: Rate at which the data in a pipeline process within a specified time.

Reliability: Requires the various systems in the data pipeline to be tolerant to faults. Therefore, a reliable pipeline has built-in auditing, validation, and logging systems that ensure data quality.

Latency: Refers to the time required for one unit of data to pass through the data pipeline. It is essentially about response time than throughput.

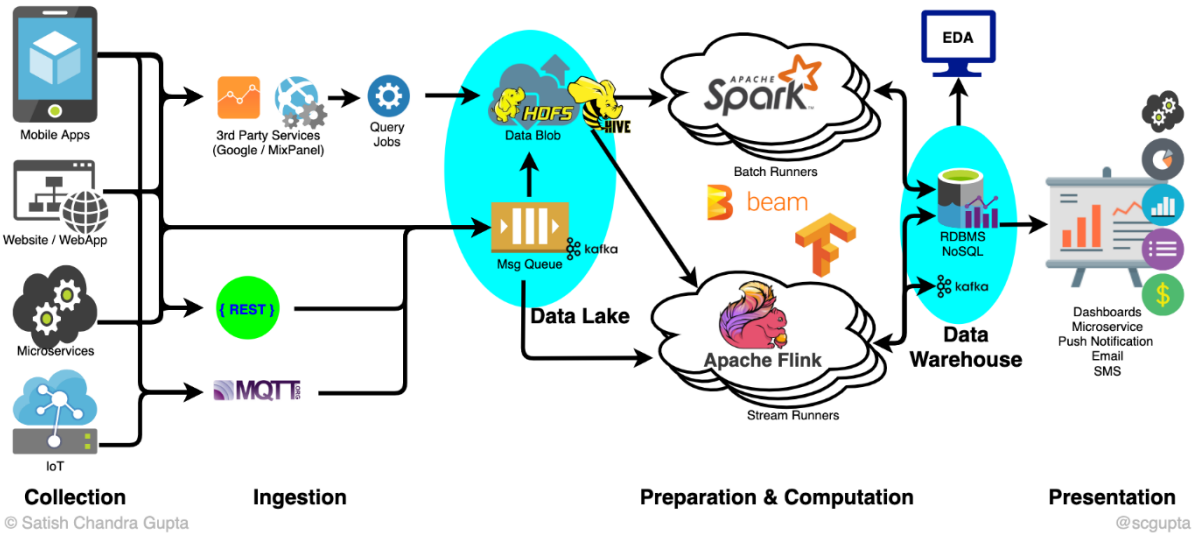
As the covid data is dynamic in nature and as every piece of information might prove vital, I am doing with Streaming Data Pipeline as it performs operations on data in motion or real-time. It enables us to swiftly sense conditions within a smaller time period from getting the data. As a result, we can enter data into the analytics tool the moment it is created and obtain prompt results.

A simple **data streaming pipeline** diagram can be shown as follows:



It can collect data from multiple sources like https://www.google.com/url?q=http://clinicaltrials.gov&sa=D&source=editors&ust=1622554599533000&usg=AFQjCNFPAqunft0EKgwalDaBT3glvRAEZA, etc., cleanse it and load the data model as mentioned in the ER model discussed in previous question.

An overall data pipeline can be thought of as follows:



A brief explanation about basic parts and processes of a Data Pipeline is as follows:

1. **Data Source:** Components of data ingestion pipeline help retrieve data from diverse sources, such as relational DBMS, APIs, Hadoop, NoSQL, cloud sources, etc.

2. **Extraction:** Some fields might have distinct elements like a zip code in an address field or a collection of numerous values. If these discrete values need to be extracted or certain field elements need to be masked, data extraction comes into play.

3. **Joins:** As part of a data pipeline architecture design, it's common for data to be joined from diverse sources. Joins specify the logic and criteria for the way data is pooled.

4. **Standardization:** Often, data might require standardization on a field-by-field basis. This is done in terms of units of measure, dates, elements, color or size, and codes relevant to industry standards.

5. **Correction:** Datasets often contain errors, such as invalid fields like a state abbreviation or zip code that no longer exists. Similarly, data may also include corrupt records that must be erased or modified in a different process. This step in the data pipeline architecture corrects the data before it is loaded into the destination system.

6. **Data Loading:** After the data is corrected and ready to be loaded, it is moved into a unified system from where it is used for analysis or reporting. The target system is can be a relational DBMS or a data warehouse.

This is a very high-level idea of how we can collect the data from multiple sources, cleanse it and load it into the model discussed in previous question.