**Big Data Project Scope:**

**Batching Analytics on Healthcare Data Using MIMIC-III**

**Objective:**

To design and implement a big data pipeline for real-time and batch processing of the MIMIC-III dataset, enabling healthcare analytics such as patient monitoring, length-of-stay prediction, and readmission risk analysis. This project focuses on integrating big data tools like Hadoop, Spark, Hive, Spark Streaming, and Flink & Kafka and Airflow.

**Title:**

"Building a Big Data Pipeline for Real-Time Healthcare Analytics with MIMIC-III"

**Learning Goals:**

1. Understand big data concepts and tools in a practical context.
2. Gain hands-on experience in batch and real-time data processing.
3. Learn to implement scalable pipelines for healthcare analytics.

**Project Requirements:**

**1. Data Preparation**

* **Docker Compose Repo :**

[**GitHub - Marcel-Jan/docker-hadoop-spark: Multi-container environment with Hadoop, Spark and Hive**](https://github.com/Marcel-Jan/docker-hadoop-spark)

* **Dataset:** MIMIC-III Clinical Database.

[MIMIC-III Clinical Database Demo v1.4](https://physionet.org/content/mimiciii-demo/1.4/)

* **Data Loading:**
  + Extract structured data such as demographics, lab results, and admission information.
* **Preprocessing:**
  + Clean data and convert it into a big data-friendly format (e.g., Parquet, Avro).
  + Store the dataset in a distributed Hadoop-based file system (HDFS).
* **Schema Design:**
  + Define schemas for tables in Hive for batch queries and real-time use cases.

**2. Big Data Pipeline Components**

**A. Hadoop for Distributed Storage**

* **Objective:** Store and manage the MIMIC-III dataset in a distributed manner.
* **Tasks:**
  + Set up Hadoop and HDFS to store large chunks of the MIMIC-III dataset.
  + Perform distributed processing using Hadoop’s MapReduce for simple analytics (e.g., calculating average patient age).

**B. Hive for Batch Processing**

* **Objective:** Perform SQL-based analysis of the MIMIC-III dataset.
* **Tasks:**
  + Create Hive tables for structured data like patient demographics and admission details.
  + Write HiveQL queries for batch analytics, such as:
    - Average length of stay per diagnosis.
    - Distribution of ICU readmissions.
    - Mortality rates by demographic groups.

**3. Deliverables**

1. **Big Data Infrastructure:**
   * Hadoop-based storage with the MIMIC-III dataset in HDFS.
2. **Hive Analytics:**
   * Hive tables and query results for batch processing.
3. Data Model:

* Comprehensive documentation including dimensions, measures, and their relationships, providing a clear overview of the data structure.

1. **Documentation:**

* User manuals and technical specifications for the ETL pipeline, including setup instructions and maintenance guidelines.

1. **Presentation:**
   * A summary of the pipeline architecture, analytics results, and project outcomes.
2. Git Repository:

* All deliverables should be organized in a private Git repository to be shared with me .

**Key Learning Areas:**

1. **Hadoop:**
   * Distributed storage and MapReduce for parallel data processing.
2. **Hive:**
   * SQL-based big data queries for structured datasets.
3. Docker :

* Containerization platform for packaging applications and their dependencies.

1. **GitHub**

* Version control platform for managing and collaborating on code using Git.
* Facilitates tracking changes, code reviews, and team collaboration.