International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 11, Issue 2, Jul–Dec 2022; 23–40

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

SCALABLE DATA ENGINEERING SOLUTIONS FOR HEALTHCARE: BEST PRACTICES WITH AIRFLOW, SNOWPARK, AND APACHE SPARK

Pramod Kumar Voola[1], Pranav Murthy[2], Ravi Kumar[3], Om Goel[4] & Prof.(Dr.) Arpit Jain[5]

[1]Burugupally Residency, Gachibowli, Hyderabad, Telangana, India,

[2]3rd Phase, Bengaluru, Karnataka, India

[3]Behind May Flower School, Patna, Bihar, India

[4]Independent Researcher, Abes Engineering College Ghaziabad, India

[5]KL University, Vijaywada, Andhra Pradesh, India

ABSTRACT

Having the capacity to handle and analyse enormous volumes of data in an efficient and effective manner is very necessary in the continually changing environment of the healthcare industry. In order to meet the ever-increasing needs for real-time data processing, sophisticated analytics, and the integration of many data sources, it is vital to have data engineering solutions that are scalable. Within the context of the healthcare industry, this article investigates the most effective methods for developing scalable data engineering solutions by using three well-known technologies: Apache Airflow, Snowpark, and Apache Spark.

The open-source workflow management technology known as Apache Airflow is an essential component in the process of orchestrating complicated data pipelines. The capabilities of this software to develop, plan, and monitor processes guarantees that data engineering activities may be automated and controlled with accuracy. In a healthcare environment, where it is essential to integrate data from a variety of sources, such as electronic health records (EHRs), wearable devices, and clinical trials, the flexibility and extensibility of Airflow make it possible to create pipelines that are both fault-tolerant and resilient. The article provides an overview of how to make use of the dynamic scheduling, task dependencies, and monitoring capabilities that are available in Airflow in order to increase operational efficiency and simplify data processing simultaneously. A big step forward in the integration and processing of data inside Snowflake's cloud data platform is represented by Snowpark, the data engineering library that Snowflake has developed specifically for this purpose. This offers a robust framework for implementing data transformations and code for data science inside a programming environment that is already known to the user. Snowpark's capability to do calculations on encrypted data assures compliance with standards such as HIPAA, which are of the utmost importance in the healthcare industry, where data privacy and security are of the utmost importance. Best practices for using Snowpark to carry out complicated data transformations, develop scalable data models, and provide support for analytics projects are discussed in this article. All of these activities are carried out while maintaining high standards of data security and governance. When it comes to analysing massive amounts of data in a timely and effective manner, Apache Spark, which is a unified analytics engine, shines. Because of its capabilities for computing in memory and its support for a wide variety of data sources, it is an excellent option for healthcare applications that need real-time analytics and batch processing. In this article, we look into the best practices for using Spark in healthcare settings. These best practices include optimising Spark tasks, utilising its machine learning library (MLlib) for predictive analytics, and connecting Spark with other data systems in order to

improve data accessibility and insights. Healthcare organisations are able to construct data engineering solutions that are scalable and efficient by combining Airflow, Snowpark, and Spark. These solutions are designed to meet the unique issues that are associated with the management of healthcare data. The purpose of this article is to demonstrate how these technologies may be coupled to develop end-to-end data engineering pipelines, increase data quality, and enable advanced analytics and decision-making processes. The study gives real examples and case studies to show it. In general, the use of these technologies and best practices helps healthcare organisations to realise the full potential of their data, drive innovation, and ultimately enhance the results for their patients. When it comes to implementing scalable data engineering solutions that are resilient, secure, and adaptable to the ever-changing requirements of the healthcare business, the purpose of this paper is to give a thorough guidance for healthcare data engineers and IT experts. Data integration, data privacy, real-time analytics, scalable data solutions, and healthcare data engineering are some of the keywords here. Other keywords are Snowpark, Apache Spark, and Apache Airflow.

KEYWORDS: *Healthcare Data Engineering, Apache Airflow, Snowpark, Apache Spark, Data Integration, Data Privacy, Real-Time Analytics, Scalable Data Solutions*

Article History

Received: *18 Sep 2022* | ***Revised:*** *19 Sep 2022* | ***Accepted:*** *20 Sep 2022*
