Chapter 4
Creating a Data Lakehouse for a South African Government-Sector Learning Control Enforcing Quality Control for Incremental Extract-Load-Transform Pipe

Dharmesh Dhabliya
https://orcid.org/0000-0002-6340-2993
Vishwakarma Institute of Information Technology, India

Jambi Ratna Raja Kumar
https://orcid.org/0000-0002-9870-7076
Genba Sopanrao Moze College of Engineering, India

Vivek Veeraiah
Sri Siddharth Institute of Technology, Sri Siddhartha Academy of Higher Education, India

Ritika Dhabliya
ResearcherConnect, India

Sukhvinder Singh Dari
https://orcid.org/0000-0002-6218-6600
Symbiosis Law School, Symbiosis International University, India

Sabyasachi Pramanik
https://orcid.org/0000-0002-9431-8751
Haldia Institute of Technology, India

Ankur Gupta
https://orcid.org/0000-0002-4651-5830
Vaish College of Engineering, India

ABSTRACT

The Durban University of Technology is now engaged in a project to create a data lake house system for a Training Authority in the South African Government sector. This system is crucial for improving the monitoring and evaluation capacities of the training authority and ensuring efficient service delivery. Ensuring the high quality of data being fed into the lakehouse is crucial, since low data quality negatively

DOI: 10.4018/979-8-3693-1582-8.ch004

impacts the effectiveness of the lakehouse system. This chapter examines quality control methods for ingestion-layer pipelines in order to present a framework for ensuring data quality. The metrics taken into account for assessing data quality were completeness, accuracy, integrity, correctness, and timeliness. The efficiency of the framework was assessed by effectively implementing it on a sample semi-structured dataset. Suggestions for future development including enhancing by integrating data from a wider range of sources and providing triggers for incremental data intake.

INTRODUCTION

In South Africa, Sector Education and Training Authorities (SETAs) are government-established entities responsible for managing skills development and training in various sectors of the economy. These entities are referred to as Government-Sector Training Authorities (GTAs), and they play a crucial role in the country's efforts to improve skills and training across many sectors. The Durban University of Technology (DUT) has partnered with a South African Government Technical Agency (GTA) to enhance the data management strategy used by the GTA for an ongoing project. This is achieved by assisting DUT students in developing supplementary talents. In order to enhance the data management capabilities of the GTA, it was recognized that a comprehensive system was required to store data and generate reports automatically. Following the discussion, the DUT team suggested using Microsoft Azure services to establish a data warehousing solution. Additional context for the project is presented by (Mthembu et al. 2024). This chapter focuses on establishing a Data Lakehouse for a Training Authority in the South African Government sector. One crucial aspect is investigating techniques to ensure the integrity of data as it moves through the system, particularly during the Incremental Extract-Load-Transform Pipelines at the Ingestion Layer employing Data Orchestration.

In the age of Big Data, where the vast amount of information poses both advantages and challenges, effectively handling and using data has become essential for organizational success. The wide range of data types, including structured, semi-structured, and unstructured forms, requires a sophisticated approach for data manipulation (Azad et al., 2020). The Extract, Load, Transform (ELT) framework is a versatile tool that is well acknowledged for its efficacy in negotiating the complexities of contemporary data settings (Singhal & Aggarwal, 2022). However, as data pipelines get larger and more complicated, guaranteeing the integrity of data quality (DQ) has become more important.

The emergence of big data has necessitated the use of Distributed Data Warehouses (DLH). Harby and Zulkernine (2022) suggested that the big data age has brought up new issues for traditional Data Warehouses (DWs). The increase in diverse data quantities caused by digital transformation presents a difficulty for traditional data warehouse solutions in businesses (Čuš & Golec, 2022; Giebler et al., 2021). Furthermore, (Barika et al. 2019) highlight the challenges faced by researchers in organizing, controlling, and implementing big data workflows, which differ significantly from typical workflows. After undergoing transformation and being placed into the data warehouse (DW), the original filtered information is no longer retained (Figueira, 2018). According to Conventional, (Nambiar and Mundra 2022), the ETL procedure is deemed inadequate for fulfilling certain data management requirements.

A Data Lake (DL) is a comprehensive storage and exploration system specifically intended to manage large amounts of varied data. It has been widely recognized as the preferred method for processing and storing various data (Begoli et al., 2021).An further research undertaken by the DUT team emphasizes