Chapter 5
Enterprise Data Lake Management in Business Intelligence and Analytics: Challenges and Research Gaps in Analytics Practices and Integration

Mohammad Daradkeh
Yarmouk University, Irbid, Jordan

ABSTRACT

The data lake has recently emerged as a scalable architecture for storing, integrating, and analyzing massive data volumes characterized by diverse data types, structures, and sources. While the data lake plays a key role in unifying business intelligence, analytics, and data mining in an enterprise, effective implementation of an enterprise-wide data lake for business intelligence and analytics integration is associated with a variety of practical challenges. In this chapter, concrete analytics projects of a globally industrial enterprise are used to identify existing practical challenges and drive requirements for enterprise data lakes. These requirements are compared with the extant literature on data lake technologies and management to identify research gaps in analytics practice. The comparison shows that there are five major research gaps: 1) unclear data modelling methods, 2) missing data lake reference architecture, 3) incomplete metadata management strategy, 4) incomplete data lake governance strategy, and 5) missing holistic implementation and integration strategy.

DOI: 10.4018/978-1-7998-5781-5.ch005

INTRODUCTION

The digital transformation towards capturing and analyzing big data opens up new ways for enterprises to optimize their processes and improve their productivity and competitive advantage. For example, the Internet of Things (IoT) applications enable the continuous collection of data directly from the production line, which in turn enable descriptive, predictive and prescriptive analytics of manufacturing and service operations (Beheshti, Benatallah, Sheng, & Schiliro, 2020; Ravat & Zhao, 2019). External data sources, such as data from mobile and social networks, can also be integrated and analyzed to gain insights that lead to a better understanding of customers' behavior, characteristics and problems. Advanced analytics approaches, such as data mining, text mining and machine learning, can generate new knowledge from a variety of sources. By applying both traditional business intelligence (BI) and data mining, collectively described as business analytics (Llave, 2018), enterprises can gain real-time insights into multidisciplinary business functions and improve overall decision making to ultimately achieve better business performance and competitive advantage (Miloslavskaya & Tolstoy, 2016).

Typically, the data used for BI and analytics applications in enterprises are heterogeneous, complex, and very large. Relevant data for BI and analytics applications can also be available in different structures and in internal and external data sources. It is not uncommon for enterprises to have multiple data silos that can only be dismantled with great effort; which poses significant challenges to traditional BI and analytics practices based on data warehouses. This is because data warehouses are traditionally not flexible enough to handle such a variety of data and usage scenarios (Beheshti et al., 2020; Inmon, 2016; Nargesian, Zhu, Miller, Pu, & Arocena, 2019). To bring all this data together and be able to extract valuable insights and knowledge from it, it can be integrated into an enterprise data lake; a scalable data repository and management platform for data exploration and analytical purposes (Gryncewicz, Sitarska-Buba, & Zygała, 2020; Llave, 2018).

In the enterprise data lake, diverse types of data, including structured, semi-structured and unstructured data, are captured, managed and analyzed using exploratory analysis and data mining without a predefined schema. For this purpose, huge amounts of data are collected from different sources (local or outside the organization) and directly ingested and stored into a data lake in its native and original state, with little or no cleansing, standardization, or transformation (Fang, 2015; Laurent, Laurent, & Madera, 2020; Llave, 2018). The native format of the data enables the transition from descriptive to predictive and prescriptive as the data management and structure can be imposed at the time of analysis, unlike traditional structured data storage where data must be mapped to a schema at ingest (Campbell, 2015; Fang, 2015). It also ensures that data analysts, data scientists, and other self-service business users have access to abundance of raw data that they can repurpose and integrate into a variety of analytics applications as needed (Fang, 2015; Larson & Chang, 2016; Llave, 2018; Tomcy & Pankaj, 2017).

While data lakes offer a variety of advantages to increase synergies and reduce the integration effort for analytics applications (Farid, Roatis, Ilyas, Hoffman, & Chu, 2016; Gorelik, 2019; Laurent et al., 2020; Llave, 2018), enterprises still encounter several challenges when building and leveraging data lakes in practice for data analysis and integration (Llave, 2018; Sitarska-Buba & Zygala, 2020). Existing literature on the concept and individual components of the enterprise data lake is vague and inconsistent. Furthermore, there are many approaches to implementing individual components of the data lake, such as data modeling and data lake governance (Farid et al., 2016; Giebler, Gröger, Hoos, Schwarz, & Mitschang, 2019; Giudice, Musarella, Sofo, & Ursino, 2019). There are also rules and concepts for roles, responsibilities, and processes in the data lake architecture, especially with regard to data security,