Data Lakes

Anjani Kumar
University of Nebraska at Omaha, USA

Parvathi Chundi
University of Nebraska at Omaha, USA

INTRODUCTION

With the advent of the digital world, data gets generated and collected for every action -- while browsing a website, purchasing items on e-commerce websites, watching videos online, etc. These data are generated in real-time and can be in diverse formats structured (relational tables or CSV files), unstructured (text files), & semi-structured (XML or log files). These ever-increasing databases create challenges in an organization where multiple departments may generate a part of the organizational data. For an organization to generate value from the data collected by different departments, these data sources must be accessible across the entire organization, merged, and analyzed in different ways for various purposes. A ***Data Lake (DL)*** is a centralized, scalable storage location where organizational data can be stored and made available widely across the entire organization for analysis purposes.

There are different requirements for different departments of a company. The Business Intelligence team may need data arranged in a specific format to compute data cubes to create reports and visualizations to answer many business questions. In contrast, the data science team may need data in a raw format to explore future trends or build predictive models.

A **data analysis task** is a process of extracting meaningful information from a massive volume of data. It can be done in various ways, such as creating reports and visualizations to answer business questions and developing a predictive model using machine learning to find patterns. There are mainly two types of data analysis: *quantitative* and *qualitative*. Even though these two tasks are conducted differently, both approaches attempt to *tell a story* from the data. Some commonalities between the two data analyses are data reduction, answering research questions, explaining variation, etc. (Hardy, 2004). A data analysis task is also defined as the accurate evaluation and full exploitation of the data obtained (Brandt & Brandt, 1998).

There are usually four steps for doing any data analysis task (Gorelik, 2019) and they are listed below.

1. **Find & Understand:** An enterprise has vast amounts of data. This massive amount of data is saved in many databases, each containing many tables and each table containing many fields or attributes. A database is the collection of interrelated data that is stored in and managed by a database management system (DBMS) (Silberschatz, Korth, & Sudarshan, 2020). In general, DBMS uses the relational or tabular format to store data and relationships among data. Data are saved in a collection of tables. Each table has multiple *columns*, also known as *attributes*. The attribute names in the table are unique. Each row in the table stores the data as a *record*.

DOI: 10.4018/978-1-7998-9220-5.ch025

With thousands of tables at an enterprise and each table containing hundreds of fields, it is difficult, if not impossible, to locate the right data sets needed for an analysis task. As a simple example, consider the data analysis task to build sales prediction models for the northeast region of the United States. The analyst should be able to locate the tables where such data are stored among the hundreds of databases in the enterprise. It becomes complicated for an analyst to find and understand the meanings of numerous attributes of these tables. To find the tables with relevant attributes, an analyst may have to manually examine each table or enlist the help of others that might have used or created that table. Therefore, the analyst must first locate the correct fields needed for the data analysis and then understand the data/attributes in existing databases.

2. **Provision:** Once correct datasets have been located, analysts will need to access this data. Acquiring access to datasets can be tedious in an enterprise. Typically, long-time employees that worked on multiple projects tend to have access to almost all the data in the enterprise, while newer employees may have nearly no access. *Provisioning* is the process of giving the proper right to access the data set so that data can be accessed at the physical (disk) level. A *metadata store* helps provide adequate access to relevant data for analysts. A metadata store contains information about all the datasets. This can be used for finding the right datasets, and then the access request to those datasets can be created. If the data is sensitive, then a de-identified version of the data, where sensitive information is replaced with randomly generated similar information, can be generated prior to granting access.
3. **Prep:** After the provisioning phase, the relevant data is obtained. It is unlikely that the data can be used directly for analysis purposes. Usually, data is not clean and does not come in the proper format. Therefore, the *data preparation* step is applied, which includes cleaning, blending, and shaping the provisioned data. As a part of cleaning, missing and mis-formatted values are fixed, and units are normalized.

As a part of shaping, a subset of relevant fields and rows may be selected. Different tables are joined to present data in a particular format for analysis. This is done by transforming, bucketizing, and aggregating data. Bucketizing is the process of converting continuous values into a discrete set of values. Blending is a technique through which data from different sources create a single, unique dataset for analysis.

4. **Analyze:** After the data preparation step, the provisioned data is in the correct format to carry out the analysis task.

Based on the above four steps, there is a real need to provide users the flexibility to search for and retrieve the data with little overhead and make the required data available in different formats depending on the type of the data analysis task. To support this functionality, the system needs to *ingest* and preserve data in its natural raw format and make the data accessible to the different end-users by converting the raw data into their desired formats. *Data ingestion* is the process of copying data from various sources to a storage platform where DL can access it. Once data is ingested, it generally is transformed and loaded into the DL storage layer for further access. Since all the data use cases are not envisaged at the time of ingestion, the raw data is preserved. For example, a data scientist might be interested in getting data in a *parquet* format, a columnar representation of data stored in an optimized way. At the same time, a business intelligence analyst might be interested in accessing data through a visualization tool such as Tableau. Having raw data gives us the flexibility to run analytics on any time range, which

Using Open-Source Software for Business, Urban, and Other Applications of Deep Neural Networks, Machine Learning, and Data Analytics Tools
Richard S. Segalland Vidhya Sankarasubbu (2022). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-28).*
www.irma-international.org/article/using-open-source-software-for-business-urban-and-other-applications-of-deep-neural-networks-machine-learning-and-data-analytics-tools/307905