


# Data Lake Architecture: A New Repository for Data Engineer

Arvind Panwar, GGSIP University, Delhi, India

 <https://orcid.org/0000-0001-9957-6365>

Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India

## ABSTRACT

Data is the biggest asset after people for businesses, and it is a new driver of the world economy. The volume of data that enterprises gather every day is growing rapidly. This kind of rapid growth of data in terms of volume, variety, and velocity is known as Big Data. Big Data is a challenge for enterprises, and the biggest challenge is how to store Big Data. In the past and some organizations currently, data warehouses are used to store Big Data. Enterprise data warehouses work on the concept of schema-on-write but Big Data analytics want data storage which works on the schema-on-read concept. To fulfill market demand, researchers are working on a new data repository system for Big Data storage known as a data lake. The data lake is defined as a data landing area for raw data from many sources. There is some confusion and questions which must be answered about data lakes. The objective of this article is to reduce the confusion and address some question about data lakes with the help of architecture.

## KEYWORDS

Big Data, Data Lake, Enterprise Data Warehouse, Hadoop

## 1. INTRODUCTION

Data is the biggest assets after people for business, and it is a new driver of the world economic and social changes for today's world. The volume of data that enterprise gathering every day is growing rapidly (Bala, Boussaid, & Alimazighi, 2017; Hefer, 2007). Every organization has its own data warehouse to store huge amount of business data. A data warehouse is designed to capture and store business data from another enterprise system for example, inventory system, supply chain management system, customer relationship management system. A data warehouse system allows business users and data analysts to drive values from data and make important decisions to grow their business.

The world is changing with speed of light so new technology has come in market for data storage, data processing, and data analysis. New technologies including streaming data, data from connected devices on internet of things, cloud computing, social media, high tech power grid, is driving a much greater volume of data (CITO research, 2014; Hortonworks, 2014). This greater volume of data is driving higher user's expectations and globalization of economics. Data generated from above-said resources is not only huge in term of volume but generate with high velocity and variety of data such as structured, unstructured and semi-structured. This kind of generated data is known as Big Data. The traditional data warehouse is not suitable to process and analyze Big Data. Now organizations

DOI: 10.4018/IJOCI.2020010104

are understanding that traditional data warehouse technologies can't match their business need to compete in the ever-growing market.

As a result, every organization is turning toward Apache Hadoop for Big Data storage and gain insights from data. Hadoop is an open-source software which is used for distributed processing and distributed storage of huge amount of data sets on computer clusters commodity hardware. Apache Hadoop provides many services like storage of data, processing of data, data access, data governance, data security, data visualization, and operations. Adoption of Hadoop in organization is growing exponentially, according to Gartner survey in mid-2015, 26% enterprises already deploying and piloting Hadoop for practice next-generation data storage and processing framework. According to survey, 12% is planning to deploy very soon and 7 to 10 percent deploy within a year.

Many organization experiences good success and growth in business with these early pursuits of mainstream Hadoop deployment in healthcare, retail, financial and e-commerce sectors. In starting Hadoop is used as tactical tools instead of strategic tool, because many opposed to replacing data warehouse. They have some questions and doubts about whether Hadoop can match their enterprise services for scalability, security, performance, and availability. But organizations know that they can't continue with data warehouse due to some challenges which come with advancement in technology.

As technology advancement enterprise data warehouse is not suitable for data storage for current market demand. Enterprise data warehouse works on the concept of schema-on-write architecture, to get data in data warehouse an extraction, transformation, and loading (ETL) process is required (Cha, Park, Kim, Pan, & Shin, 2018; Khine & Wang, 2018). With this architecture, organization design a data model and prepare an analytic plan before loading data. In other words, organization must know in starting, before loading data, how they are planning to use that data, and this is very limiting. Big data analytics want data storage who works on schema-on-read concept in which data is stored in raw format as data generated or in other words, there is no need to prepare an analytic plan before loading data, and no need to know ahead of time how they plan to use that data.

Enterprise data warehouse store data that has been modeled or structured but Big Data analytics in the market need storage who store raw data and store all kind of data such as structured, unstructured, semi-structured and quasi-structured data at one place. To fulfill market demand researchers, work on new data repository system for Big Data storage known as data lake. The idea of data lake is to enhance enterprise data warehouse environment (Mrozek, Dabek, & Małysiak-Mrozek, 2019; Nogueira, Romdhane, & Darmont, 2018). The data lake is defined as a data landing area for the raw data from many and always increases number of data source in organization. Data from data lake can be transformed and distributed to the downstream system as they required. Now it's clear that data lake supports Big Data initiatives and data lake approach can reduce data silos (Sawadogo, Scholly, Favre, & Ferey, 2019; Shepherd, Kesa, Cooper, Onema, & Kovacs, 2018; Singh, 2019). The data lake is the requirement of the industry for data storage but there are some confusion and question which must be answered about data lake. For example, how to design & deploy data lake? How to govern and secure data lake? What kind of data that can be managed in data lake? Why do organizations need data lake? The objective of this survey paper is to reduce the confusion and addressing the above mention question with the help of data lake architecture.

In this paper, author focuses on Big Data and data lake and try to reduce confusion about data lake. The author describes how data lake is better than data warehouse in current scenario for Big Data analytics. This paper is divided into different section. In next section author describe Big Data concept of this paper. In section 3 author explains about data lake concept and difference between data warehouse and data lake. In section 4 author give a reference architecture of bid data lake. In last section 5 author give conclusion of the paper and future research direction.

11 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/data-lake-architecture/243678](http://www.igi-global.com/article/data-lake-architecture/243678)

## Related Content

---

### Dynamic Particle Swarm Optimization with Any Irregular Initial Small-World Topology

Shuangxin Wang, Guibin Tian, Dingli Yuand Yijiang Lin (2015). *International Journal of Swarm Intelligence Research* (pp. 1-23).

[www.irma-international.org/article/dynamic-particle-swarm-optimization-with-any-irregular-initial-small-world-topology/137085](http://www.irma-international.org/article/dynamic-particle-swarm-optimization-with-any-irregular-initial-small-world-topology/137085)

### Creative Accelerated Problem Solving (CAPS) for Advancing Business Performance

Cookie M. Govender (2021). *Handbook of Research on Using Global Collective Intelligence and Creativity to Solve Wicked Problems* (pp. 84-109).

[www.irma-international.org/chapter/creative-accelerated-problem-solving-caps-for-advancing-business-performance/266781](http://www.irma-international.org/chapter/creative-accelerated-problem-solving-caps-for-advancing-business-performance/266781)

### An Evolutionary Algorithm Based Approach for Business Process Multi-Criteria Optimization

Nadir Mahammedand Sidi Mohamed Benslimane (2017). *International Journal of Organizational and Collective Intelligence* (pp. 34-53).

[www.irma-international.org/article/an-evolutionary-algorithm-based-approach-for-business-process-multi-criteria-optimization/180311](http://www.irma-international.org/article/an-evolutionary-algorithm-based-approach-for-business-process-multi-criteria-optimization/180311)

### Spectrum Access Issues and Security in Cognitive Radio Network

Mamata Rath (2019). *International Journal of Organizational and Collective Intelligence* (pp. 31-44).

[www.irma-international.org/article/spectrum-access-issues-and-security-in-cognitive-radio-network/223189](http://www.irma-international.org/article/spectrum-access-issues-and-security-in-cognitive-radio-network/223189)

### Particle Swarm Optimization Algorithms Applied to Antenna and Microwave Design Problems

Sotirios K. Goudos, Zaharias D. Zaharissand Konstantinos B. Baltzis (2013). *Swarm Intelligence for Electric and Electronic Engineering* (pp. 100-126).

[www.irma-international.org/chapter/particle-swarm-optimization-algorithms-applied/72825](http://www.irma-international.org/chapter/particle-swarm-optimization-algorithms-applied/72825)