DATA LAKE

A Paradigm Shift in the Next Generation Reservoirs

Data lake typically uses low-cost commodity servers in a scale-out architecture where servers can be added as needed to increase processing power and data capacity. In comparison, data warehouse can't be scaled cost-efficiently to process the growing data volume. Data lake provides fast access to targeted data for valuable business insights in dynamically changing scenario.

SAVITA BHATNAGAR
Technical Director
savita.bhatnagar@nic.in

Edited by
P. LENIN

Data Lake is emerging as a new paradigm to store variety of structured, semi-structured, and unstructured data in its native format, without much prior processing, as required in conventional data warehouses for data analytics. It is a schema-less repository where data is classified, organized or analyzed only when it is accessed.

BIG DATA

Big data refers to voluminous amount of structured or unstructured data collected from multiple sources in diverse formats. Gartner defines big data as "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". The rapid advancement in information technology is leading to the exponential growth of data in diverse formats from various sources such as social media, press media, blogs and Internet of Things (IoT) etc. According to International Data Corporation (IDC), the data is expected to be around 44 ZB in 2020. Data scientists are continuously working on new and innovative ways to manage big data for advanced analytics.

DATA LAKE

Governments and business organizations in today's digital world are dependent on how the data is stored, managed, processed and protected for better decision–making. The majority of this data is unstructured and difficult to manage or process efficiently in traditional way. Data lake is emerging as a solution to store, manage and analyze large and quickly arriving volumes of unprocessed structured, semi-structured, and unstructured data.

Data lake is a schema-less massively scalable storage repository that holds vast amount of raw data in its native format. Data is not preprocessed before storing in the repository, as the value and analysis requirements are not clear at the outset. It is classified, organised or analyzed only when it is accessed. Advance dynamic analytical applications are used to access datasets and analyze the data. This helps the organization to address business and operational challenges which are difficult to address using traditional data warehouse technologies.

Data lake typically uses low-cost commodity servers in a scale-out architecture where servers can be added as needed to increase processing power and data capacity. In comparison, data warehouse can't be scaled cost-efficiently to process the growing data volume. Data lake provides fast access to targeted data for valuable business insights in dynamically changing scenario. The key aspects of data lake are:

COLLECT EVERYTHING

Data lake is a central repository which contains all data, even though the scope of data or its use is not known.

SCALABILITY AT LOW COST

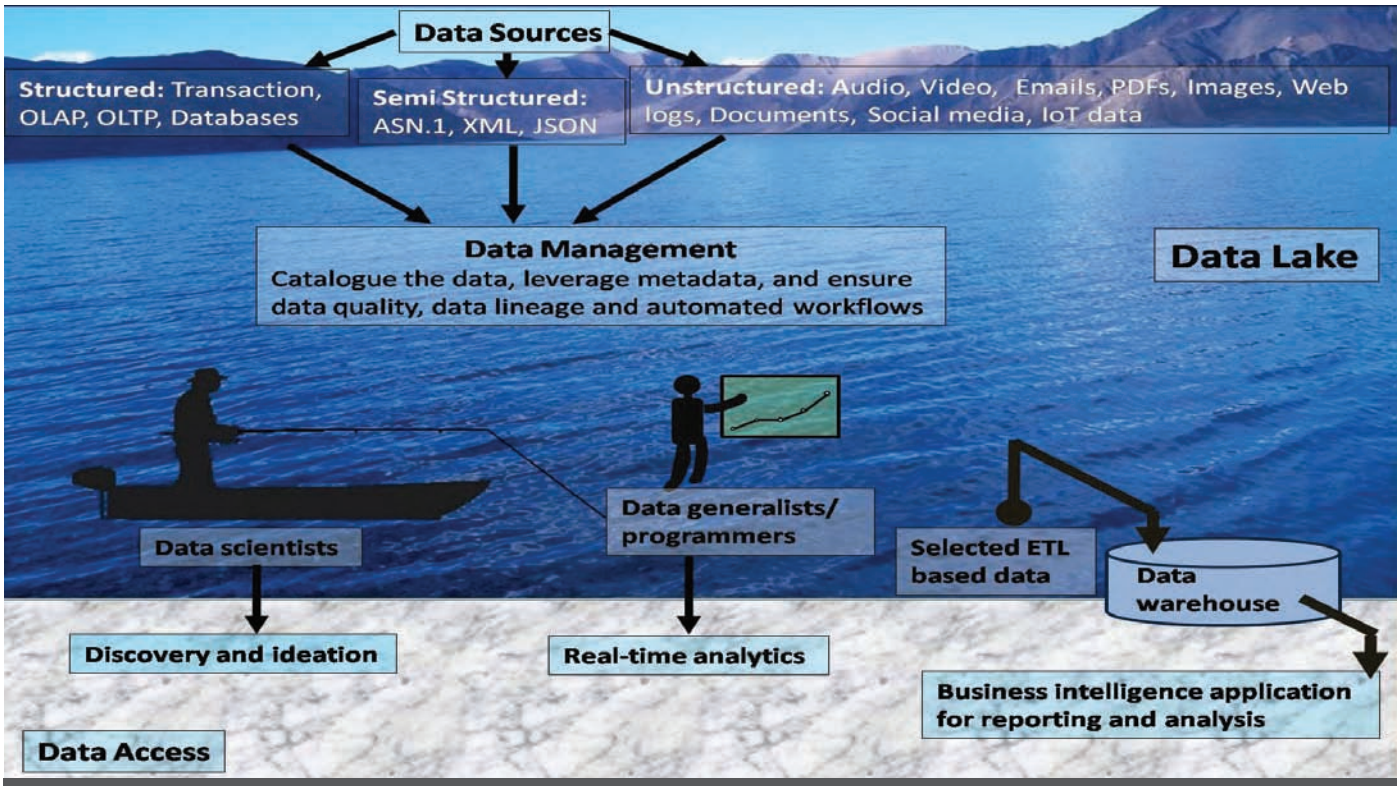
Huge amounts of data can be stored in data lake without much prior processing. Only the data that is going to be analyzed needs to go through processing step, thereby reducing the data storage cost.

HETEROGENEOUS DATA IN THE SAME REPOSITORY

Multi structured data from diverse set of sources such as logs, XML, multimedia, sensor data, binary, social data, chat and people data can be stored in data lake. It handles all kind of data regardless of its source or format.

SCHEMA ON READ

There is no need to create a schema before capturing the data. Schema is created only when reading the data.



Components of Data Lake Architecture

ADVANCED ANALYTICS

Data lake removes the need for data modeling at the time of ingestion, which can be done at the time of consuming. It offers unmatched flexibility to ask any business domain questions and to seek valuable insightful information. It enables data scientists to explore new data sources and analytic techniques for better results and predictions.

FLEXIBLE ACCESS

Data lake enables multiple data access patterns across a shared infrastructure like batch, interactive, online, search, in-memory and other processing engines. Data scientists have the flexibility to change their hypothesis at any given time without worrying about data availability.

DATA SOURCES

The data lake allows collection of data for future needs before it is possible to know what those needs are. Data ingested in lake from various sources, including structured data from files and databases (OLTP, OLAP etc.), semi-structured data (ASN.1, XML, JSON) and unstructured data (emails, documents, pdfs, images, audio, video etc). A data lake provides massive storage for any type of information that comes from multiple resources and arrives in multiple formats. Data flows into data lakes on the basis of real time, incremental, batch or one time copy.

DATA MANAGEMENT

Data lake management platform ingest and manage large volumes of diverse data sets in the data lake. It allows cataloging the data, leveraging metadata, and supports the ongoing process of ensuring data quality, data lineage, and automating workflows. Each element in data lake is assigned a unique identifier and tagged with a set of extended metadata.

DATA ACCESS

Data access services allow outside tools or applications to access data stored in the lake, regardless of the format or type of persistence like analytics programs, business intelligence resources and a range of other applications. There are many ways to access data from the whole sets to individual objects. File transfers, APIs, SQL queries, even search are all possible mechanisms for accessing data stored in the lake. Selected data which has gone through the process of Extraction and Transformation can be Loaded (ETL) in data warehouse.

FUTURE OPPORTUNITIES

Data access and management in digital world is becoming a critical priority, especially in the backdrop of big data that is being created in variety of formats from diverse sources. Data lake is emerging as a new approach to store and manage big data and uses it, as and when required, for advanced data analytics. Governments across the world are also dealing with massive amount of structured and unstructured data, which may require creation of their own data lakes. This will enable governments to access diverse data sets for real-time, data-driven decision-making and achieving new insight.

For further information, please contact:

SAVITA BHATNAGAR
Technical Director
NIC State Centre,11th Floor
New Administrative Building, Mantralaya
Madam Cama Road, Mumbai, MAHARASHTRA

Email: savita.bhatnagar@nic.in
Phone: 022-24139151