Architecture of Data Lake

Author, Ajit Singh

A Data Science Foundation White Paper

April 2019

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: contact@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

ABSTRACT

Data can be traced from various consumer sources. Managing data is one of the most serious challenges faced by organizations today. Organizations are adopting the data lake models because lakes provide raw data that users can use for data experimentation and advanced analytics. A data lake could be a merging point of new and historic data, thereby drawing correlations across all data using advanced analytics. A data lake can support the self-service data practices. This can tap undiscovered business value from various new as well as existing data sources.
My paper will present the overview of data lake, benefits and it's architecture along with the opportunities laid down by data lake and advanced analytics, as well as, the challenges in integrating, mining and analyzing the data collected from these sources. It goes over the important characteristics of the data lake architecture and Data and Analytics as a Service (DAaaS) model.

Keywords: Data lake, Overview, benefits, architecture, underlying models, layers of architecture

1. **INTRODUCTION**

A data lake is a centralized data repository that can store a multitude of data ranging from structured or semi-structured data to completely unstructured data. Data lake provides a scalable storage to handle a growing amount of data and provides agility to deliver insights faster. A data lake can store securely any type of data regardless of volume or format with an unlimited capability to scale and provides a faster way to analyze datasets than traditional methods.

A data lake provides fluid data management fulfilling the requirements of an industry as they try to rapidly analyze huge volumes of data from a wide range of formats and extensive sources in real-time.

A data lake has flat architecture to store data and schema-on-read access across huge amounts of information that can be accessed rapidly. The lake resides in a Hadoop system mostly in the original structure with no content integration or modification of the base data. This helps skilled data scientists to draw insights on data patterns, disease trends, data abuse, insurance fraud risk, cost, and improved outcomes and engagement and many more. A data lake gives structure to an entity by pulling out data from all possible sources into a legitimate and meaningful assimilation. Adopting data lake, means developing a unified data model, explicitly working around the existing system without impacting the business applications, alongside solving specific business problems.

However, with every opportunity comes a challenge. The concept of "Data Lake" is challenging, the attributing reasons being Entities have several linkages across the enterprise infrastructure and functionality. This leads to non-existence of a singular independent model for entities.

It contains all data, both structured and unstructured, which enterprise practices might not support or have the techniques to support.

It enables users across different units of enterprise to process, explore and augment data based on

the terms of their specific business models. Various implementations might have multiple access practices and storage construct for all entities

Technology should be able to let organizations acquire, store, combine, and enrich huge volumes of unstructured and structured data in raw format and have the potential to perform analytics on these huge data in an iterative way. Data lake may not be a complete shift but rather an additional method to aid the existing methods like big data, data warehouse etc. to mine all of the scattered data across a multitude of sources opening new gateway to new insights.

2. **Key Benefits Of Data Lake**

1. Scalability: The Hadoop is a framework that helps in the balanced processing of huge data sets across clusters of systems using simple models. It scales up from single server to thousands, offering local computation and storage at each node.

   Hadoop supports huge clusters maintaining a constant price per execution bereft of scaling. To accommodate more one just has to plug in a new cluster. Hadoop runs the code close to storage getting massive data sets processed faster. Hadoop enables data storage from disparate sources like multimedia, binary, XML and so on.

2. High-velocity Data: The data lake uses tools like Kafka, Flume, Scribe, and Chukwa to acquire high-velocity data and queue it efficiently. Further they try to integrate with large volumes of historical data.

3. Structure: The data lake presents a unique arena where structure like metadata, speech tagging etc. can be applied on varied datasets in the same storage with intrinsic detail. This enables to process the combinatorial data in advanced analytic scope.

4. Storage: The data lake provides iterative and immediate access to the raw data without pre-modelling. This offers flexibility to ask questions and seek enhances analytical insights.

5. Schema: The data lake is schema less write and schema-based read in the data storage front. This helps to develop up to date patterns from the data to grasp applicable intelligent insights without being dependent on the data.

3. **Architecure of Data Lake**

Data lake architecture should be flexible and organization specific. It relies around a comprehensive understanding of the technical requirements with sound business skills to customize and integrate the architecture. Industries would prefer to build the data lake customized

to their need in terms of the business, processes and systems.

An evolved way to build a data lake would be to build an enterprise model taking few factors into consideration like, organization's information systems and the data ownership. It might take effort but provides flexibility, control, data definition clarity and partition of entities in an organization. Data lake's self-dependent mechanisms to create process cycle to serve enterprise data, help them in consuming applications.

1. **Data Governance and Security Layer**
2. **Metadata Layer**
3. **Information Lifecycle Management Layer**

Tiers are abstractions for a similar functionality grouped together for the ease of understanding. Data flows sequentially through each tier. While the data moves from tier to tier, the layers do their bit of processing on the moving data. The following are the three tiers:

1. **Intake Tier**
2. **Management Tier**
3. **Consumption Tier**

One major architecture defining data lake architecture is the Lambda architecture pattern. This architecture makes the data lake fault tolerant, data immutable and helps in re-computation.

The CAP theorem, also named as Brewer's theorem, states that a distributed data store cannot simultaneously provide more than two out of the following three:

1. **Consistency**
2. **Availability**
3. **Partition tolerance.**

The data lake with Lambda Architecture's aid, works with the CAP theorem on a contextual basis. The three major contributions of the CAP theorem are Consistency, Availability and Partition tolerance. Usually availability is chosen over consistency because consistency can be achieved eventually. If not most data goes with approximations.

DAaaS (Data Analytics-as-a-Service) is a protractible platform. It uses a cloud-based delivery model. It provides a wide range of tools to select from for data analytics that can be designed by the user to process large amounts of data effectively. Enterprise data is ingested into the platform. Further the data is processed by analytics applications. This could provide business insight using advanced analytical algorithms and machine learning

As per researchers, experts and data enthusiasts, the "Data Lake" to "a successful Data and

Analytics" transformation needs the following:

DAaaS Strategy Service Definition: Our Informationists leverage define the catalog of services to be provided by the DAaaS platform, including data onboarding, data cleansing, data transformation, datapedias, analytic tool libraries, and others.

DAaaS Architecture: We help our clients achieve a target-state DAaaS architecture, including architecting the environment, selecting components, defining engineering processes, and designing user interfaces.

DAaaS PoC: We design and execute Proofs-of-Concept (PoC) to demonstrate the viability of the DAaaS approach. Key capabilities of the DAaaS platform are built/demonstrated using leading-edge bases and other selected tools.

DAaaS Operating Model Design and Rollout: We customize our DAaaS operating models to meet the inpidual client's processes, organizational structure, rules, and governance. This includes establishing DAaaS chargeback models, consumption tracking, and reporting mechanisms.

DAaaS Platform Capability Build-Out: We provide the expertise to conduct an iterative build-out of all platform capabilities, including design, development and integration, testing, data loading, metadata and catalog population, and rollout.

4. **Conclusion**

Data lakes with advanced analytics are reshaping the way enterprises work. Future with data lakes looks very promising. System developers are immersed in vigorous R&D for such technology advancement for better analysis and detail oriented search. It could be useful for industries by providing better efficiency and outcomes.

To be at an advantage, industry will have to use the power of data lake driven processes and systems. If fathomed intuitively, it could change the way services is being delivered.

Presently, data lake practices are governed by Hadoop predominantly. Hadoop has become the major tool for assimilating and pulling out insights from combinatorial unstructured data present in Hadoop and enterprise data assets, running algorithms in batch mode using the MapReduce paradigm. Hadoop, with the existing enterprise data assets such as data in mainframes and data warehouses. Languages such as Pig, Java Map Reduce, SQL variants, R, Hadoop, Apache Spark, and Python are being increasingly used for data munging, data integration, data cleansing, and running distributed analytics algorithms.

There is more to consider with details including: big data architecture for accessible Data Lake infrastructure, data lake functionality, solving data accessibility and integration at enterprise level, data flows in the data lake, and many more. With these numerous queries, there still is resources

to tap and a lot to gain for the enterprise. Using the data lake architecture to derive cost efficient, life-changing insights from the huge mass of data nullifies the concern regarding going further with the ice-berg hidden under the ocean.

REFERENCES

- https://tdwi.org/articles/2017/03/29/executive-summary-data-lakes.aspx
- Data Lake Development with Big Data by Beulah Salome Purra, Pradeep Pasupuleti http://www.datasciencecentral.com/profiles/blogs/9-key-benefits-of-data-lake
- https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation