Mphasis
The Next Applied

Never lose data by implementing Enterprise Data Lake

Whitepaper by
Srinivas Challa
Associate Vice President
Analytics Practice

What is an Enterprise Data Lake

A Data Lake, as the name says, is a single place where I can dump any kind of data irrespective of the format. There is no limitation on the types of files a data lake can store. Yes, that is its capacity. However, enterprises typically capture all analytically useful data onto one single infrastructure. From there, we can apply schema-on-read approach using analytical applications. It also depends on the enterprise vision and charter on what kind of data has to be stored on the lake.

How Data Lake is different from EDW?

The difference is obvious and quite simple. An Enterprise Data Warehouse stores only structured data, which is implemented on RDBMS. It follows 'schema-on-write', meaning the data type is checked while loading a table. A Data Lake, in contrast, can store all types of data without restrictions. Not only that, the data in EDW is cleansed, massaged, and transformed as per the EDW principles. You can store anything and everything in the data lake.

You may question - like what am I going to get from all types of data? You are at least consolidating them at one single place and getting analytics and intelligence out of it. You may or may not need to convert them to structured format, depending on the analytical tools.

> Enterprise Data Warehouse and Data Lake have distinct use. While the former can be used to store structured data, the latter is home to every kind of data.

Well, my idea is NOT to get rid of Enterprise Data Warehouse. It is still needed and it has its well defined role and importance. A data lake can be used to compliment an Enterprise Data Warehouse. For instance, it could be used for discovering the hidden or unutilized enterprise data and see what kind of data is best for your next EDW phase. I mean discover the hidden data mines in your organization. Many enterprises are concerned about the cost to store the huge volumes of data and ROI from them.

Why Data Lake – Lots of data getting damaged before it turns to Information

The simple answer is huge volume of data and a place to store more variety of data. There is lots of data in the enterprise (in the world also for that matter) that is not utilized nor analyzed properly and that tremendous data is being damaged before it turns to Information. The examples range from sensors, scanners, web clicks, weather info … to social media posts.

> The prospect of deriving the information out of data is huge. Let's start with storing the entire data in a Data Lake before it dies.

Why can't I read tweets about my company and make use of it for campaigns, sales, bug fixing or simply know about my company/product's talk in the public?

Can I have a system that watches the server log files and raise service tickets automatically when the resource utilization crosses the threshold limit?

Can I analyze electricity consumption of my city and take necessary actions to balance the power consumption and reduce the peak hour utilizations?

Okay, let us talk about our enterprise internal transactional data. How many times your data architect or DBA said "NO" for the availability of certain enterprise data? You may not want to hear the words "Not Available" again. Sometimes the data doesn't have the granularity that you wanted, sometimes the data is not available beyond certain history, and sometimes it is archived to save some hardware resources.

Data Lake with Hadoop – Benefits

The word "Data Lake" became familiar with Big Data. In my view, the term big data is something with huge volume or variety that it is difficult to be handled by the existing technologies. When we hear the word Big Data, the very next thing that comes to our mind is the baby elephant Hadoop. Agree that it was the name of the toy elephant, but it is not small any more.

Hadoop comes with many features that solve your big data problems. Let us take a look at the top ones.

1. **Open Source** – No license fee and save many $.
2. **Distributed and Parallel** – Hadoop is distributed on several commodity machines and runs in parallel.
3. **Cost Effective** – You don't need big servers with large RAM, CPU Power etc. A group of commodity machines are fine with Hadoop. This group is called Hadoop Cluster in Hadoop world.
4. **Brings the program to the data** – Let us say you have 1 GB file and it is distributed among 10 machines in the Hadoop cluster. While executing any application that uses this data, Hadoop does not bring the entire data to the main server. Hadoop copies the program/ application logic to these 10 machines, runs the logic on them in parallel and then consolidates.
5. **Flexible** – It is easy to add new machines to the Hadoop cluster without bringing the cluster down.

Well, you may get lot of questions/concerns like how can I rely on the commodity machines? What if a machine fails? Hadoop 2 Architecture has solid reliable answers for these concerns and they are transparent (open source) on how they are doing these.

> Big Data storage needs strong architecture. Thus, Hadoop can be a distinct choice for Data Lake.

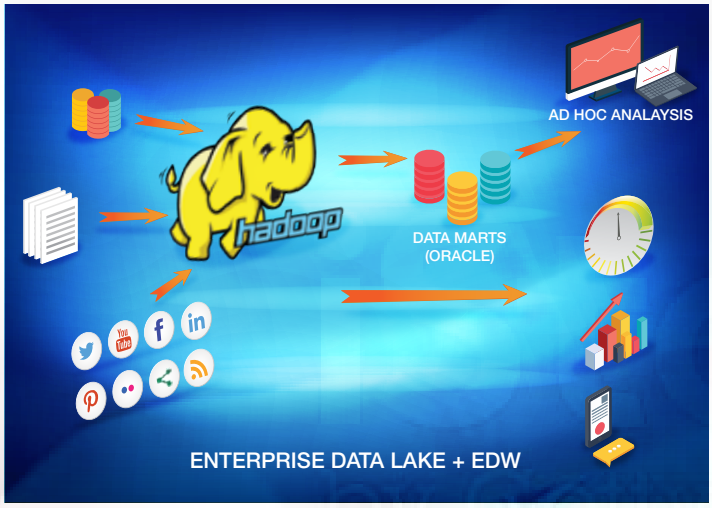
Hadoop Supported File Systems

The default Hadoop file system is HDFS (Hadoop Data File System). You can also mount Hadoop on other popular distributed file systems such as Amazon S3 (Simple Storage Service), IBM GPFS (General Parallel File System) and WASB (Windows Azure Storage Blob).

Industry Use Case – Hadoop + EDW (Right Technology at the Right Place)

A leading healthcare service provider company in the US, serving more than 25 states of United States, is using the Data Lake for storing the complete enterprise data from all data sources of the company.

The data sources include both internal sources (enterprise applications, members, claims, and call data) and external sources (CMS, Hospitals, smart pill bottles and social media). The data lake is the one stop shopper for the entire enterprise data.



The data from all internal applications and external sources is first stored in the Enterprise Data Lake, in raw format. So, the granularity and originality is preserved. The data is then massaged, transformed and aggregated in the data lake (as per the business needs) and fed to the EDW. A copy of the transformed data is also stored in the Enterprise Data Lake.

In this case study, we have seen the Enterprise Data Lake acting as pre-processer for the EDW. The EDW stores the data that is needed for the business intelligence only, however, the Enterprise Data Lake stores everything. Users query the Hadoop Data Lake.

> A data lake can store pre-processed data also in its native form. Its architecture keeps the unused and unstructured data secured for future needs.

Benefits of this Architecture:

- Data is never lost. It is preserved in the raw format first.
- Users can find the analyzable data that is ignored in the past.
- Storing the aggregated/transformed/integrated data in the data lake helps to serve the single version of truth to the enterprise.
- Enterprise Data Lake is also accessible for ad hoc queries (limited to advanced users). It helps the business to prioritize new subject areas and data for the next phase of EDW.

Careful

Even the experienced Hadoop Data Lake users say that a successful implementation requires a strong architecture and disciplined data governance policies, without which the Data Lake systems can go out of control. As I mentioned in the beginning, you can store anything and everything on the Data Lake. However, if you don't have proper governance and guidelines for the lake, though it is cheap, it can be a dump yard.

Srinivas Challa

Associate Vice President, Analytics Practice

Srini Challa has 16 years of overall IT experience in designing and developing analytics solutions using Big Data and DW BI methodologies. He played different roles in his IT journey, including Data Architect, BI Architect, ETL Architect, DW BI Solutions architect and Big Data Solutions Architect. Srini is currently working for Mphasis Analytics Practice, leading the Big Data COE and responsible for Next Labs delivery. Srini is passionate on data management and analytics.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized (C = X2C2™ = 1) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655
Fax: +1 212 683 1690

USA
226 Airport Parkway
San Jose
California, 95110
USA

UK
88 Wood Street
London EC2V 7RS, UK
Tel.: +44 20 8528 1000
Fax: +44 20 8528 1001

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000
Fax: +91 80 6695 9942

www.mphasis.com

VAS 8/07/16 US LETTER BASIL4/20