DIGILYTICS

Data Lakes

A New Approach to Managing Data

What is a data lake?

A **data lake** is a large storage area of structured and unstructured data. With a relational database a predefined schema must be defined. However, with a data lake the schema is defined on a need basis only, James Dixon who coined the term data lake described it as being "*If you think of a DataMart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."*

How does a data lake create value for a business?

Is a data lake just another buzz word? Or does it actually provide, tangible measurable value for an organization?

At Digilytics, we don't believe the question has been completely answered one way or the other yet. However, a number of organizations are now actively considering this. Here are some examples of value benefits that data lakes have created for businesses.

- Cross-channel analysis of customer interactions and behaviors data from multiple sources provided a more complete view of the customer
- Integrating loan origination, loan servicing, CRM and Finance data provided improved lending with lower cost of originating each loan
- Correlating brand equity data, consumer panel data, and campaign data improved lead conversion from targeted marketing campaigns

Risks of Data Lakes

Having a data lake does not mean that it does not have to properly managed. If not managed properly a data lake can transform itself to a data swamp, i.e. a large storage area of data that cannot be leveraged for analytics and thereby any useful business insights. More recently, there is also an emerging debate on the security profile of data that can be maintained in a data lake since it has enterprise-wide access.

Features of Data lakes

Historically, a data lake was usually implemented to address some of the key pain points of using an enterprise data warehouse

- Reconciling conflicting data needs
- Providing real-time access
- Assembling data from multiple sources
- Supporting ad hoc analysis

Large enterprises with two or more silos of data might want to think of implementing a data lake for data source integration as well as real-time access of all data for all divisions of an organization.

Increasingly, Data Lakes are being used to

- To capture and store raw data at scale for a low cost

- To store many types of data in the same repository: like integration of silos in a large organization
- To perform transformations on the data
- To define the structure of the data at the time it is used, referred to as schema on read: Not necessary to define the schema before the data need to be used.
- To perform new types of data processing: Different data lends the
- To perform single subject analytics based on very specific use cases

Examples of the kinds of data stored in Data lakes:

- Enterprise data
- Clickstream data
- Server logs
- Social media data
- Geolocation coordinates
- Machine and sensor data

How is it different from a Data Mart and a Data Warehouse?

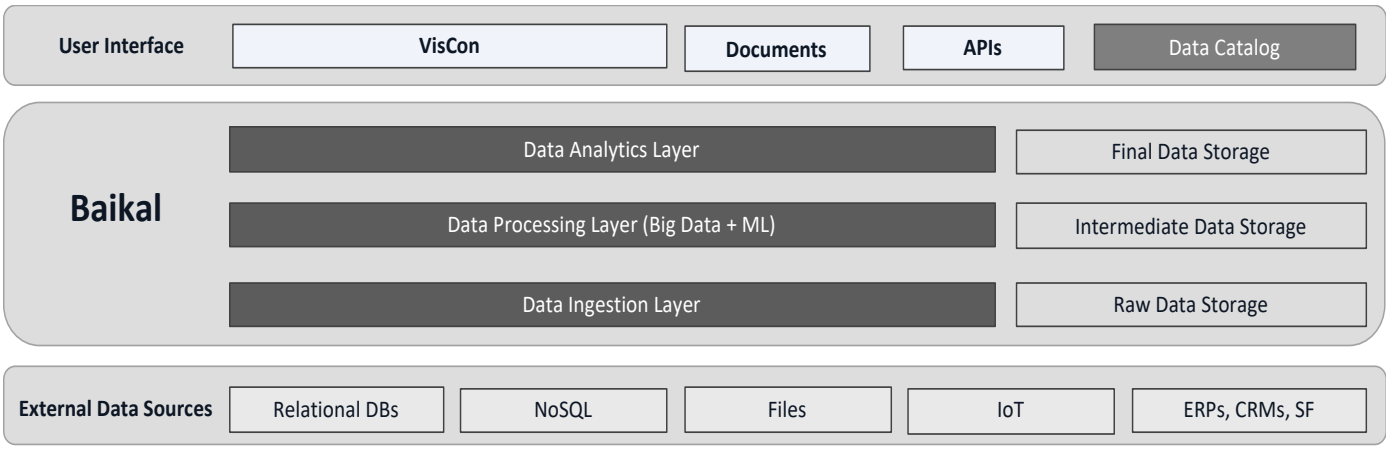
Is it better to keep an enterprise data warehouse (EDW), or is it worth investing in data lake? Here we provide a table for comparison of a data lake and an EDW.

High level Comparison between Data Warehouse and Data Lakes

Dimension	Enterprise Data Warehouse	Data lake
Workload	• Batch processing	• Batch processing at scale
Schema	• Schema on write	• Schema on read
Scale	• Scale to large data volumes at moderate cost	• Scale to large data volumes at low cost
Access methods	• SQL/BI tools	• SQL-like systems
Benefits	• Fast response times • Consistent performance • High concurrency • Easy consumption of data • Rationalization of data from multiple sources into a single enterprise view • Clean, safe, and secure data • Cross-functional analysis • Transform once, use many	• Execution on thousands of servers with excellent quality • Parallelization • Support for higher level programming frameworks • Changes economic model for storing high volumes of data

Dimension	Enterprise Data Warehouse	Data lake
SQL	• ANSI SQL, ACID compliant	• Flexible programming, evolving SQL
Data	• Cleansed	• Raw
Access	• Seeks	• Scans
Complexity	• Complex joins	• Complex processing
Cost/ Efficiency	• Efficient use of CPU/IO	• Low cost of storage processing

The Digilytics data lake architecture allows easy ingestion and management of data:



Benefits of data lake over a data warehouse

Following are the benefits of implementing a data lake over a conventional data warehouse:

1. **Volume, Variety and Velocity**
   The real strength of the data lake is that it does a good job addressing some of those Big Data challenges like volume, variety, and velocity. It's infinitely scalable, handles structured or unstructured data, and is designed for rapid data ingestion. Because it's schema-on-read, we don't have to understand the format of the data until we're ready to read it back out. We can write that data very quickly before we have that understanding and without risk of write errors.

2. **Not Designed for Big Data**
   Big Data challenge is one of the major sticking points when it comes to the data warehouse. Its schema-on-write is not really optimized for that variety, velocity and volume of data. Disruptors like hybrid source systems (cloud and on-prem), and high data volume, variety and velocity coming from big data scenarios can limit the effectiveness of the data warehouse.

3. **Limited Exploration**
   Cost and warehouse model can restrict capturing data of unknown value, which can limit exploration. You might only pull certain amounts of data where you had known reporting requirements. There may be a whole other set of data that you are not bringing in that users might want to explore.

4. **Accessible Data**
   Landing the data in the data lake makes it easy for users to open copies of that data (or subsets of that data) to different user groups. Whether that's for self-service or data science, they can control access to that data.

5. **Latency**
   With a data warehouse, you have to wait for each business process component to be built to get value from that data. Although there is a lot of value in having an analytical model, it also takes time to have your development team support getting data into that model, typically through the development of ETL processes. If you have data that's not in your data warehouse today, and users want to report on it, then there can be additional latency to get that data into the warehouse.

6. **Low Cost Storage**
   A data lake is also low cost (relatively speaking). That allows users not to worry as much about the types of data that they are storing. Users may not know the analytical value around their data yet, but with data lake storage, the focus is on inexpensively hanging on to it for some future point where they might be able to find value for it.

What technologies are used in the Digilytics data lake?

The Digilytics data lake leverages the Microsoft Azure stack. While constructing the architecture of a data lake, the following considerations should be kept in mind:

Dimensions	Description
Insights	Ability to analyse all the data with varying QoS (real-time, interactive, batch) to generate insights for business decisioning
Action	Ability to integrate insights with business decisioning systems to build data-driven applications
Unified data management	Ability to manage the data lifecycle, access policy definition, and master data management and reference data management services
Unified operations	Ability to monitor, configure, and manage the whole data lake from a single operations environment
Storage	Ability to store all (structured and unstructured) data cost efficiently in the Business Data Lake
Ingestion	Ability to bring data from multiple sources across all timelines with varying Quality of Service (QoS)
Distillation	Ability to take data from the storage tier and convert it to structured data for easier analysis by downstream applications
Processing	Ability to run analytical algorithms and user queries with varying QoS(real-time, interactive, batch) to generate structured data for easier analysis by downstream applications

Stages of Data Lake implementation

Implementing a data lake is a journey that has many stages:

1. Phase 1: Define the high-level requirements of the data lake and develop the high-level architecture

2. Phase 2: Identify and implement use-cases that address pressing business problems that can provide clear, quick, and measurable business value opportunities

3. Phase 3: Move from a reactionary approach to a proactive approach where we begin to use data from social networks and internet of things and other data for big data and analytics projects

4. Continuously optimize the data lake.

At Digilytics, we have expertise in advising clients interested in implementing Data Lakes.

About Digilytics AI

At Digilytics™, we aim to drive business value leveraging our platform. In an ever-crowded world of clever technology solutions looking for a problem to solve, our solutions start with a keen understanding of what creates and what destroys value in your business. Founded in 2014, by Arindom Basu, the leadership of Digilytics™ is deeply rooted in leveraging disruptive technology to drive profitable business growth. With over 50 years of combined experience in technology-enabled change, the Digilytics™ leadership is focused on building a values-first firm that will stand the test of time. The leadership strongly believes in the ethos of enabling intelligence across the organization. Digilytics™ is headquartered in London, with presence across India.

About the Authors

Arindom Basu, Arindom.basu@digilytics.ai

CEO and Founder at Digilytics AI. Experienced technology disruptor with a demonstrated history of working in the financial services and consumer industry sectors. Entrepreneur with a degree in computer sciences and management, Arindom has spent almost 30 years witnessing technology evolution through crisis times.

Nalin Suri, Nalin.suri@digilytics.ai

Product Manager & business Consultant with over 7 years of experience in IT consulting, product strategy and product implementation for B2B & B2C web & mobile products across Accenture, Virtusa and Appster

Get in touch with us

London Office:
85 Gresham Street
London – EC2V7NQ
Call on: +44 208 947 0137
ask@digilytics.ai

Gurugram Office:
408 Centrum Plaza
Sector 54, Gurugram
Call on: +91 124 467 3910
ask@digilytics.ai

Bibliography

- Jacobsohn, M & Delurey, M (2013). How the Data Lake Works, Booz Allen Hamilton. Retrieved from https://www.capgemini.com/resource-file-access/resource/pdf/pivotal-business-data-lake-technical_brochure_web.pdf

- Capgemini (2013) The Principles of the Business Data Lake. Retrieved from https://www.uk.capgemini.com/resource-file-access/resource/pdf/the_principles_of_the_business_data_lake_2013-12-02_v07_web.pdf

- CITO (2014 April). Putting the Data Lake to Work: A Guide to Best Practices. Retrieved from https://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf

- Capgemini (2013). The Technology of the Business Data Lake. Retrieved from https://www.capgemini.com/resource-file-access/resource/pdf/pivotal-business-data-lake-technical_brochure_web.pdf

- Stein, B. & Morrison, A (2014). The enterprise data lake: Better integration and deeper analytics. Retrieved from https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

- O'Brien, John (2015 February): The Definitive Guide to the Data Lake. Retrieved from http://www.teradata.co.uk/Resources/White-Papers/The-Definitive-Guide-to-the-Data-Lake/?LangType=2057&LangSelect=true