

A White Paper By

* Ashwil JS
* Jithin Dev
* Richa Iyer

Published: Dec 2016

Contents

[List of Figures 3](#_Toc470598237)

[Abstract 3](#_Toc470598238)

[Introduction 4](#_Toc470598239)

[The Challenges of Traditional Data Warehousing 5](#_Toc470598240)

[Can Business Data Lake provide solutions? 6](#_Toc470598241)

[The Data Lake approach to handle data 7](#_Toc470598242)

[Architecture of Data Lake 9](#_Toc470598243)

[Architecture Comparison of Traditional DW and Data Lake 12](#_Toc470598244)

[Benefits of Business Data Lake 14](#_Toc470598245)

[Strategies & Considerations before diving into Data Lake 15](#_Toc470598246)

[Hybrid Approach Vs Stand Alone 19](#_Toc470598247)

[Future of Data Lake 21](#_Toc470598248)

[Conclusion 24](#_Toc470598249)

[References 25](#_Toc470598250)

# List of Figures

*Figure 1: Data Pipeline*

*Figure 2: Sample Architecture*

*Figure 3: Unified Data Management Tier*

*Figure 4: Data Lake Vs Traditional BW Architecture Comparison*

*Figure 5: Comparison of Data Lake and Traditional DW capabilities*

*Figure 6: Physical Layers of Modern Data Architecture - Example*

*Figure 7: Independent solutions?*

*Figure 8: A hybrid solution to fill the gaps*

# 

# Abstract

*“Data scientists as well as large enterprises have been pondering over ways to fetch relevant information from the ever increasing and diverse data inputs. Traditional Data warehousing seems to be stretched to its limits and finds itself trying to match pace with the latest demands in data analytics. The concept of Data Lake is emerging as a popular way to organize and build the next generation of systems to tackle big data challenges, but there are concerns and questions for large enterprises to implement it, as Data Lake is still in its evolving stage. This whitepaper attempts to explore the potential of Data Lake to handle the future enterprise data requirements as well as its capability to challenge the strengths of traditional data warehouses. Further, the study looks into how traditional data warehousing can possibly evolve, and, whether a hybrid architecture of Traditional DW and Data Lake will prove to be a better option to handle the latest data analytics demands, than stand-alone solutions.”*

# Introduction

The success or failure of a business enterprise in the digital age depends on its use of information – and specifically on its ability to deliver insights at the point of action based on all the available data. This means that ease of access to information, and the quality of that information, are more critical than ever.

*The traditional data warehouse has served businesses well for many years, but new trends are causing it to break in four different ways:*

* *Data growth*
* *Fast query expectations from users*
* *Non-relational/unstructured data*
* *Cloud-born data*

Organizations therefore need new enterprise data management platforms that create maximum value from data. The platforms must be able to combine existing data resources with new types of data, both structured and unstructured – including data associated with social media, documents, and the Internet of Things. They must make the combined data available for both fast real-time analytics in operations and strategic predictive analytics.

Data Lake is an evolving concept in data warehousing which shows capability to handle latest data trends. The following study looks into the features, current capabilities and evolving trends of, Data Lake and Traditional Data Warehouse (DW), and throws light into the aspects that should be prioritized, when considering Data Lake for business analytics.

But, will Data Lake replace traditional data warehouse or co-exist to handle future data scenarios? Can traditional warehouse evolve to manage latest data trends? Is a hybrid structure of Traditional DW and Data Lake a better solution? This white paper discusses the possible solutions to the above questions.

# The Challenges of Traditional Data Warehousing

The traditional data warehouse/business intelligence approach has done a great job of simplifying data access and reporting, as well as combining data from many sources, in order to answer all of the questions an organization may have.

But it’s impossible to anticipate every question a business might ask and every report they might need. Metrics change from year to year, month to month and sometimes even day to day. In addition, there is a flood of new data types. Information from the web, social media, servers, sensors, documents, comments and devices has caused an explosion in the volume of data that organizations are trying to understand.

Traditional Business Intelligence (BI) systems provide various levels and kinds of analyses on structured data but they are not designed to handle *unstructured data*. For these systems Big Data brings big problems because the data that flows in may be either structured or unstructured. That makes them hugely limited when it comes to delivering Big Data benefits.

Further problems come with the need for *near real-time analysis*. This requires the ability to handle and process high velocity data in near-real time - a major challenge for the traditional BI implementation methods, which have data latency built into their architecture. Solutions have been developed to circumvent these issues and bring in as much data as feasible in near real-time, but these create their own problems - not least the issue of high storage volumes and costs.

The emergence of Big Data calls for a radically new approach to data management. Organizations now need near real-time analysis on structured and unstructured data. Traditional BI approaches that call for building EDWs and data marts are unable to keep up.

*Most traditional DW and BI implementations follow either a Top-Down or a Bottom-Up approach to set up the EDW and Data Marts.*

* A top-down approach is excellent as a solution to the “single source of the truth” problem, but it can fail in practice due to the long implementation cycle and a relational structure that is not friendly for business analysis on the fly.
* The bottom-up approach, while very flexible for business analysis, struggles to maintain a “single source of truth” because data redundancy is possible across data marts. Eventually, the model just becomes an integration of fragmented data marts.

# Can Business Data Lake provide solutions?

*The term data lake was coined by Pentaho CTO James Dixon. He describes a data mart (a subset of a data warehouse) as akin to a bottle of water, “cleansed, packaged and structured for easy consumption” while a data lake is more like a body of water in its natural state. Data flows from the streams (the source systems) to the lake. Users have access to the lake to examine, take samples or dive in.*

*A Business Data Lake can be defined as a data repository that can store and handle massive amounts of structured, semi-structured and unstructured data in its raw form in low cost commodity storage as it arrives. It provides the ability to perform Line of Business-specific business analyses yet present a global enterprise view of the Business. Metadata information is maintained for traceability, history and future data refinement needs.*

Let us look into the different types of business input data:

**Structured Data**

The data which can be co-related with the relationship keys, in a geeky word, RDBMS data. It concerns all data which can be stored in database SQL in table with rows and columns. They have relational key and can be easily mapped into pre-designed fields. Today, those data are the most processed in development and the simplest way to manage information. But structured data represent only 5 to 10% of all data. So let’s introduce semi structured data.

**Semi Structured Data**

The structured data which does not conform to formal structure of data models in context of relationships is semi-structured data. Examples could be XML, JSON, some NoSQL databases like MongoDB which store the data natively in JSON. This again represents another 5% of the total available data.

If the data has any organizational structure (a known schema) or carries a tag (like XML extensible markup language used for documents on the web) then it is somewhat easier to organize and analyze, and because it is more accessible for analysis may make it more valuable.

**Unstructured Data**

Data having no structure at all, falls into this category and it represents whopping 85-90% in share. It often includes text and multimedia content.

Examples could be sensor data (huge in percentage), social media streams, images, videos, mobile data, e-mail messages, word processing documents, audio files, presentations, webpages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered unstructured, because the data they contain doesn’t fit neatly in a database.

Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.

Here are some examples of machine-generated unstructured data:

1. Satellite images: This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
2. Scientific data: This includes seismic imagery, atmospheric data, and high energy physics.
3. Photographs and video: This includes security, surveillance, and traffic video.
4. Radar or sonar data: This includes vehicular, meteorological, and oceanographic seismic profiles.

The following list shows a few examples of human-generated unstructured data:

1. Text internal to your company: Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
2. Social media data: This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
3. Mobile data: This includes data such as text messages and location information.
4. Website content: This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

# The Data Lake approach to handle data

Data Lake can accept, store and process data that falls in structured, semi-structured or unstructured data type. Let us look into how Data Lake handles different data types. We would discuss the structural and technical aspects.

**All data, regardless of form, is collected into the Persistent layer of the Data Lake**

* Data from all internal and external source systems – including structured, semi-structured and unstructured data, as well as streaming sources – is gathered in a single Persistent layer in the data lake.
* Not just data that is in use today but data that may be used someday and even data that may never be used at all
* Data is also kept for all time so organizations can go back to any point in time to do analysis.

This approach becomes possible because the hardware for a data lake usually differs greatly from that used for a data warehouse. Commodity, off-the-shelf servers combined with cheap storage make scaling a data lake to terabytes and petabytes fairly economical.

The data lake approach embraces the non-traditional data types. Data lakes store all data, regardless of source and structure. Data is kept in its raw form and only transformed when it is ready for use. This approach is known as *“Schema on Read” vs. the “Schema on Write”* approach used in the data warehouse.

**Analysts and data scientists help shape and Curate the data for business use**

The Hadoop data lake provides the opportunity to create an Active Archive to store additional historical data and make it available for query for extended analytics use cases.

**Self-service analysts continue to refine the curated data into an Operational layer for broader use**

In many business cases an alternative ELT (Extract-Load-Transform) is preferable because the data lake lends itself to loading data prior to transformation. All transformation and integration is done in the layers of the data lake. Regardless of the methodology, it is important to choose appropriate tools that can be automated and audited.

**Data Lakes Support All Users**

* In most organizations, 80 percent or more of users are “operational”. They want to get their reports, see their key performance metrics or slice the same set of data in a spreadsheet every day.
* The next 10 percent or so do more analysis on the data. They use the data warehouse as a source but often go back to source systems to get data that is not included in the warehouse and sometimes bring in data from outside the organization.
* Finally, the remaining users do deep analysis. These users include the Data Scientists and they may use advanced analytic tools and capabilities like statistical analysis and predictive modeling.

The data lake approach supports all of these users equally well. The data scientists can go to the lake and work with the very large and varied data sets they need while other users make use of more structured views of the data provided for their use.

**Data Lakes Adapt Easily to Changes**

Many business questions can’t wait for the data warehouse team to adapt their system for answers. This ever-increasing need for faster answers has given rise to the concept of self-service business intelligence.

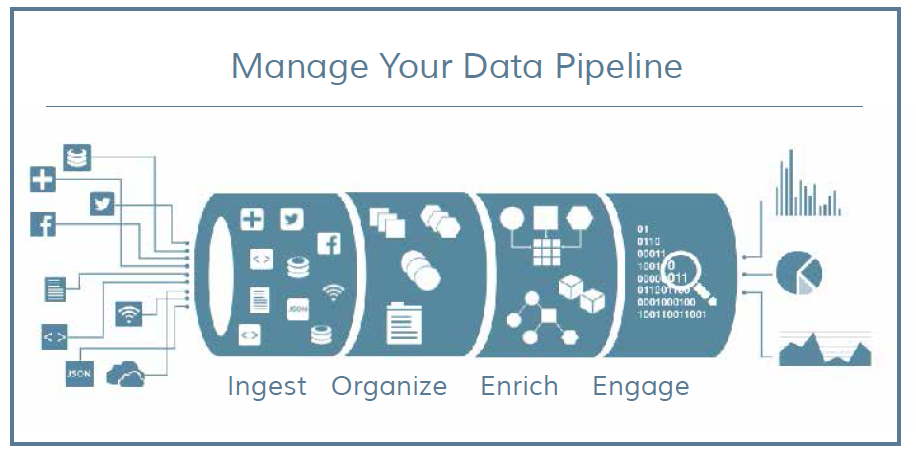
In the data lake on the other hand, since all data is stored in its raw form and is always accessible to someone who needs to use it, users are empowered to go beyond the structure of the warehouse to explore data in novel ways and answer their questions at their pace.

**Data Lakes Provide Faster Insights**

This last difference is really the result of the other four. Because a data lake contains all data and data types, because it enables users to access data before it has been transformed, cleansed and structured, it enables users to get to their results faster than the traditional data warehouse approach.

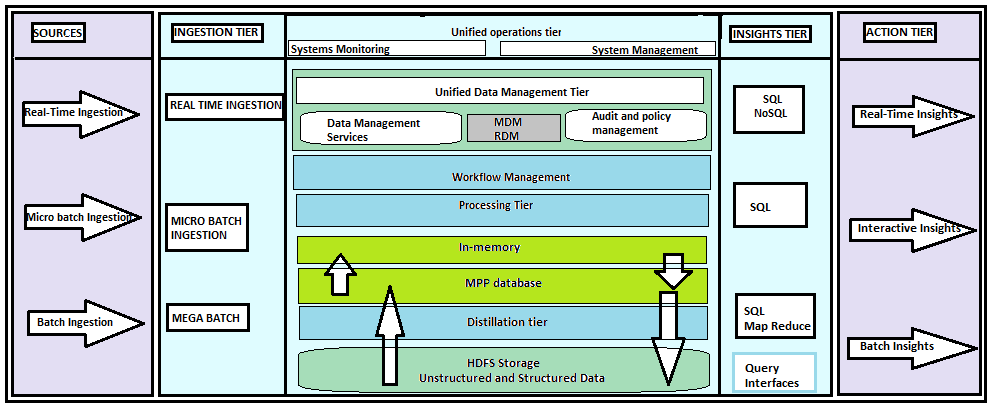
# Architecture of Data Lake

Because it processes diverse forms of data, the handling stages are different from that of Traditional warehousing. Let us try to understand the steps involved. Below diagram shows an easy to understand process flow.



#### Figure 1: Data Pipeline

The above processes are detailed in below diagram using a sample architecture of a Data Lake product package – Pivotal Business Data Lake



#### Figure 2: Sample Architecture of a Data Lake product package

The figure above shows the key tiers of a Business Data Lake. Data flows from left to right. The tiers on the left depict the data sources, while the tiers on the right depict the integration points where insights from the system are consumed.

Let us understand what each stage is meant for:

**Data ingestion**

One of the principal differences between a data lake and a traditional EDW approach is the way data is ingested. Rather than performing heavyweight transformations on data to make it conform to a canonical data model, the data can be ingested into the data lake in its native form.

* *Ingesting real-time data (“streaming”):* Real-time data needs to be collected from devices or applications as it is generated, one event at a time. Much critical streaming enterprise data is revenue bearing – for example, credit card transactions – so data quality, reliability and performance are vital.
* *Ingesting batches of data:* Custom processing can be executed on batches while the data is still being ingested, before it is stored on Hadoop HDFS. The processing can be as simple as transformations and lookup, or as complex as machine learning, scoring or address cleanup.
* *Bulk data ingest:* Often, large amounts of data need to be moved from one platform to another. A key to success in bulk data transfer is maximizing network bandwidth without impacting other applications that share the same network resources.

**Data Distillation and Processing**

Data distillation and processing are related but separate topics in a data lake environment. *Distillation* is about refining data and adding structure or value to it. *Processing* is about triggering events and executing business logic within the data tier.

**Insights**

To gain insight from your Business Data Lake, you can provide access through a standard SQL interface: this way, you can use existing query tools on big data. For example, *HAWQ* component is a full SQL query engine for big data, with a distributed cost-based query optimizer that has been refined in the Greenplum database to maximize performance for big data queries.

**Taking action on data**

Tools like, Spring Tool Suite helps to build new big- data driven applications rapidly which can then be integrated with business decision systems. In addition, integration components such as *Spring XD and* *RabbitMQ* enable to integrate the insights from data into existing business applications across the enterprise.

Now, let us look into the physical components of Data Lake architecture:

**Data storage tier**

The current explosion of both structured and unstructured data demands a cost- effective, reliable storage mechanism. The *Hadoop Distributed File System (HDFS2)* has emerged as the predominant solution, providing a low-cost landing zone for all data that is at rest in the system. One of the key principles of Hadoop is the ability to store data “as is” and distill it to add the necessary structure as needed. This “schema-on-read” principle eliminates the need for heavy extract transform load (ETL) processing of data as it is deposited into the system.

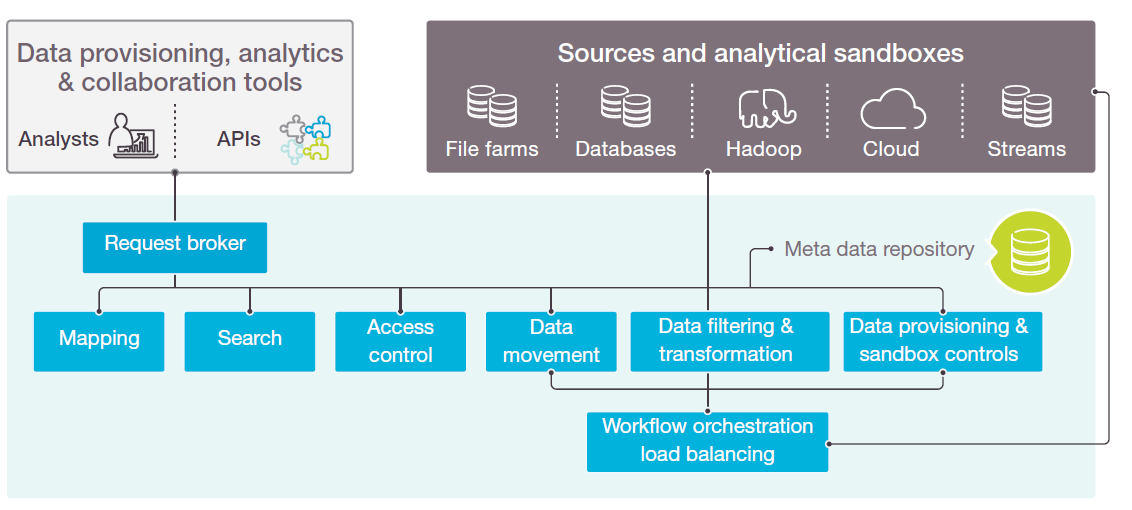
*Support for real-time responses:* Many systems need to react to data in real time. For these systems, the latency of writing the data to disk introduces too much delay. Examples include the new class of location-aware mobile applications, or applications that have to respond to events from machine sensors. For these systems, an in-memory data solution provides the ability to collect and respond to data with very low latency while it is in motion, and to persist the data on HDFS when at rest.

**Distillation tier**

With the Data Lake approach, the process of ingesting, distilling, processing, and acting upon the data does not rely on a pre-ordained canonical schema. Instead, raw data can be ingested and stored in the system in its native form until it is needed. The schema and structure are added as the raw data is distilled and processed for action using MapReduce jobs.

**Unified Data management tier**

To manage and provide access to all the data that is collected in the Business Data Lake, authorized data workers can access data sets through a self-service portal that allows them to look through a metadata catalog of the data in the system and create a single view of data from across the company. Workflows in the system allow users to lease sandboxes of data and start complex analytics processes.



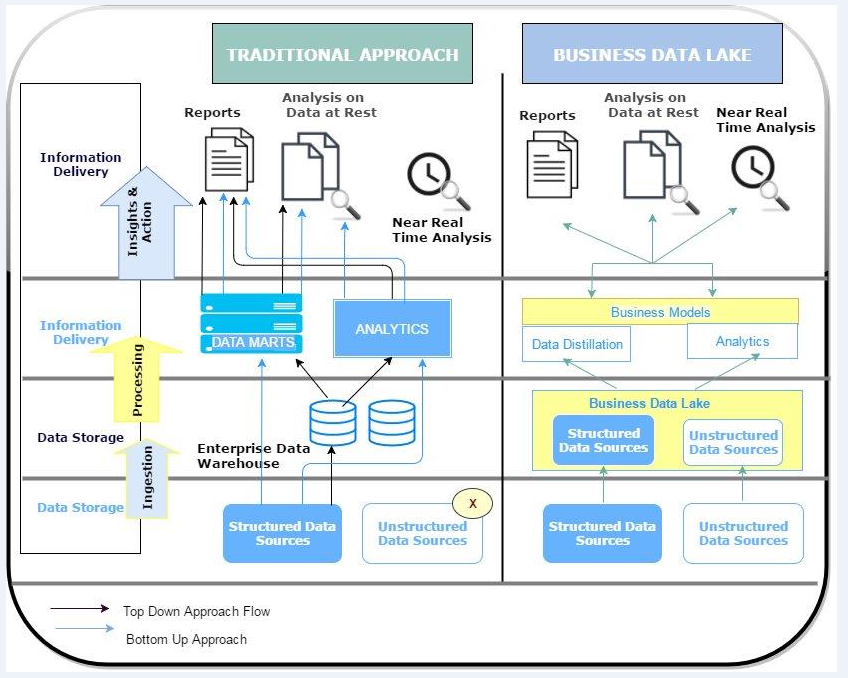
#### Figure 3: Unified Data Management Tier

**Insights Tier**

The insights can be accessed through a variety of interfaces from the Insights tier. In addition to Hadoop query interfaces like Hive or Pig, SQL – the lingua franca of the data world – can be used. Interactive analytics tools and graphical dashboards enable business users to join data across different data sets and draw insights. Real-time insights can also generate external events – for example they can trigger actions in areas like fraud detection, or send alerts to mobile applications.

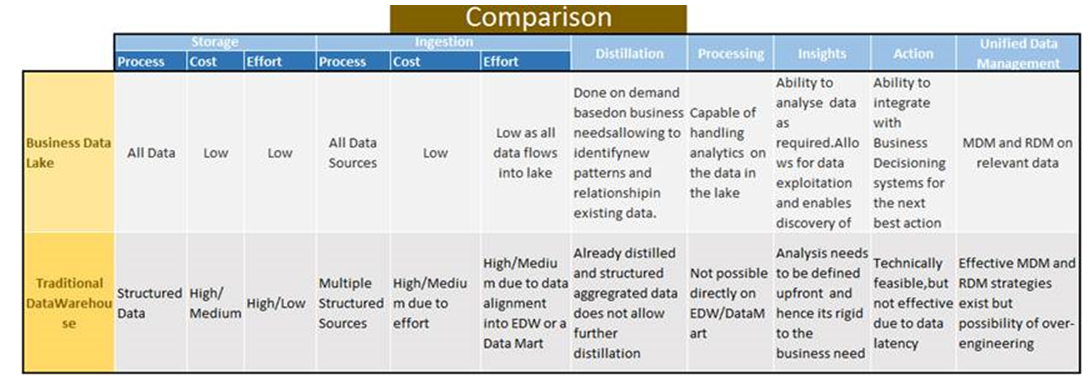
# Architecture Comparison of Traditional DW and Data Lake

Now that we have seen capabilities of both Traditional DW and Data Lake, the differences can be summarized as shown in below diagram and comparison table



#### Figure 4: Data Lake Vs Traditional BW Architecture Comparison

To simplify the above diagram let us look at the following table to do the comparison.

Figure 5: Comparison of Data Lake and Traditional DW capabilities

# Benefits of Business Data Lake

Organizations are particularly interested in the relationship between enterprise BI and data discovery. Leveraging the data lake enables data warehouse staging areas to store the detailed, raw transaction history of operational systems that can be extracted and transformed into business activities, measures and metrics.

Enterprise BI methodologies also benefit when quick data exploration for discovery, profiling and feasibility is required. A key advantage that data lakes hold over other repositories is that raw, derived and aggregated data is able to serve multiple application types.

In addition to meeting the needs of big data applications, the lake supports the same raw information for statistical functions and machine-learning algorithms. As a result, the data lake enables an enterprise architecture to support BI, discovery and data science in a unified way.

The benefits can be summarized as below:

* A Business Data Lake is a storage area for all data sources. Data can be pulled/pushed directly from the data sources into the Storage Area. All data in raw form are available in one place.
* Limitations on the data volumes and storage cost are significantly reduced through the use of commodity hardware.
* Once all data is brought into the Lake, users can pull relevant data for analysis. They can analyze and derive new insights from the data without knowing its initial structure.
* APIs that search the data structures in the Business Data Lake and provide the metadata information are currently being created. These APIs play a key role in deriving new insights from ad hoc data analysis.
* As new data sources get added to the environment, they can simply be loaded into the Business Data Lake and a data refinement/enrichment process created, based on the business need.
* The main drawback of creating a data model up-front is eliminated. Traditional data modelling, which is done up-front, fails in a Big Data environment for two reasons:
* The nature of the incoming data and the limitation on the analysis that it allows - The Business Data Lake overcomes these two limitations by providing a loosely coupled architecture that enables flexibility of analysis.
* Different business users with different needs can view the same data from different dimensions.
* Based on repetitive requirements, relevant subject areas that are used frequently for standard / canned reports can be loaded into the data warehouse in a dimensional form and the rest of the data can continue to reside inside the Business Data Lake for analytics on need.
* A data governance framework can be built on top of the Business Data Lake for relevant enterprise data. This framework can be extended to additional data based on requirements.
* The Business Data Lake meets local business requirements as well as enterprise wide needs from the same data store. The enterprise view of the data can be considered as another local view.
* Being able to move data across from the sources and turn it around quickly to derive business outcomes is key to the success of a Business Data Lake, an area where traditional BI implementations fail to meet business needs.

# Strategies & Considerations before diving into Data Lake

Data lake usage should be part of an enterprise IT application and data strategy that is architecturally sound and goes beyond the discussion of a basic commodity infrastructure, data science or machine-learning analytics. While the data lake is new to the industry’s vernacular, it is a solution that allows businesses to have conversations with a common lexicon that drive clarity and reduce the risk that data lake initiatives will become data swamps.

Let us look into the strategies an organization can adopt for a successful Data Lake implementation.

* **Start with the use case.** Have use cases in mind before constructing a data lake. They could be existing ones or any problem that your business wants to solve but currently can’t.
* **Involve all facets of the business.** Data lakes are resources for the entire organisation, not just IT. Therefore, all interested parties should be involved in planning data lake projects. Data lakes are central to your organisations data architecture, and cannot be implemented in isolation. A data lake project should involve business leaders, IT, storage leaders, and end users.
* **Analytics should drive business benefits.** The business value of a data lake has little to do with the underlying technologies selected. Business value is derived from the analytics skills you can apply to the lake.
* **Use of multiple tools and products.** Customise your lake based on your existing and available technology stacks.
* **Domain specification.** Data lakes must be specific to cater to industry’s unique needs. Make sure that IT intervention is not necessary to enable users to obtain data when they need it. A user interface that allows keyword, faceted and graphical search will be necessary.
* **Automated metadata management.** Automated and mandatory metadata management is a must to avoid the lake becoming a swamp. Attributes like data lineage, data quality, and usage history are key to usability.
* **Configurable ingestion workflows.** New sources of external information will be available continuously. Have an easy, secure and trackable content ingestion workflow mechanism that can rapidly add new information into the data lake.
* **Integrate with the existing environment.** Many businesses have existing enterprise data management systems. The data lake must be integrated and be able to support these to avoid replacing or damaging the existing environment.
* **Allow the data lake to mature gradually.** With the data lake, users can take what is relevant and leave the rest. Individual business domains can mature independently and gradually. Perfect data classification is not required. Data maturity results as a natural outgrowth of the ongoing user interaction and feedback at the metadata management layer—interaction that continually refines the lake and enhances discovery.

Informed dialogues within the organization followed by appropriate actions will allow the lake to realize its full potential by profoundly enhancing enterprise data. In addition, with the data lake’s enterprise-wide integration of operational and analytic applications, businesses are assured of efficient journeys to—and successful arrivals at—their data destinations.

The considerations for a Data Lake project must include the criteria listed here:

**Rethink Data for the Long Term**

Achieving critical mass within any data project, including data lakes, relies on several key strategies:

* Consider data lake retention after the current process has consumed the raw data. Weigh the potential reusability in future applications and discovery while recognizing that some future needs will be unknown.
* Think about acquiring new data as it is identified so more information will be readily available for future projects.
* Invest in a growing data lake to maximize compute and memory resources, and reduce data duplication and movement across the enterprise.
* Understanding the five-year goal of the cluster will allow you to plan strategies for cluster management and code & metadata organization.

**Come up with a Business Benefit Priority List**

What business value will the cluster generate? Identify what objectives are tied to the core/essential business vs. those that are value-add (e.g. for a bank, fraud detection vs. twitter sentiment analysis). Stating the business uses along with the priority will guide everything from cluster configuration (e.g. for YARN queue configuration) to ecosystem tool choices.

**Have an Architectural Oversight**

What technologies, frameworks, and tools need to be evaluated based on use cases prioritized by business and considering skills in the organization? Architecture review processes and gating processes for new systems can help maintain the Data Lake’s architectural integrity.

**Decide your Security Strategy**

Understanding data from creation to consumption is critical to data-centric security. That requires the organization to:

* Define access rights for what will be brought into the lake, and by whom.
* Determine the point at which information might need to be encrypted, and to what degree. This will dictate how security technologies work together.
* Scrutinize and understand data usage patterns to recognize how to properly secure and encrypt information in the data lake.
* Optimize security controls and set the proper level of encryption to avoid unnecessary performance degradation.

**What should be the I/O & Memory Model?**

Will the cluster be used for running statistical models or for ETL? Memory and I/O fundamentals should be reviewed and modeled out for future growth. Some Data Lakes serve as an MPP engine whereas some function as a large database, so it is important to understand the hardware and compute needs of your Data Lake.

**Skill set to take Data Lake Forward**

What skills exist within the firm that can be leveraged? Understanding your current strengths will guide the ecosystem tool selection in the near term and will help guide human resource planning. Data scientists and engineers can provide the necessary expertise required to make the data lake a successful data and analytics tool. You may choose to work with data consultants or in house staff. The key is in hiring technically adept facilitators.

**Data Governance Strategy**

What policies (e.g. lifecycle management policies such as retention, archiving, purging), audit/review procedures and authorities/roles are required to meet the business needs?

* Establish a common understanding for how managers and stakeholders will work with the data to mitigate doubts and second-guessing.
* Extend current governance policies to focus on data ingestion and evaluating internal and external sources, third-party information and specialized user data sets.
* Define key technology requirements, including access controls, mobility and security.

**Clarity on Operations Plan**

How are cluster resources, workloads and users managed? What are the procedures for making requests or investigating issues? Understand the controls of your Data Lake environment to guide how your processes and procedures can be leveraged or need to change.

**Clear Communication Plan**

How are cluster-wide changes communicated? How are requests triaged? How does a developer onboarding ramp up quickly? By asking these questions, you may find that distribution lists and possibly, new roles and responsibilities, are required in the organization.

**Disaster Recovery Plan**

What are the plans should the data center become unavailable? A DR plan is part of most enterprise systems, but in the early phases of Hadoop adoption DR is often an afterthought. Run through disaster scenarios early in new use case implementations and update the DR plan as required.

**Special Considerations:**

Data lakes can be influenced by proximity, compliance and security. The concept of proximity is related to minimizing data movement across networks. If the majority of data ingestion is cloud-based public information, organizations might want to split the data lake into a cloud-public side and an internal-enterprise side.

Compliance requirements might force companies to segregate private data that has security requirements on a portion of the data lake that is maintained on-premises. Whether the data is in the cloud or on premise, discovery environments can be used to evaluate new or external data that is later incorporated into the data lake.

# Hybrid Approach Vs Stand Alone

Traditional data warehouses are limited to structured data. Whereas, the data lake can contain any type of data: clickstream, machine-generated, social media, and external data, and even audio, video, and text.

However, the organizations that are using both an enterprise data warehouse and a data lake often face a different problem – because they are in fact creating a distributed form of analytics. Results from analysing video, audio, unstructured text, etc., are often found in the data lake, while the results from applying analytics to structured data are either found in the analytical modelling tool environment, or written back to the data warehouse for reporting and further action.

The issue with these two separate environments is that the most value is obtained when these separate sets of modelling results are integrated. In reality, in most cases, the models will produce better results when they are run over an integrated dataset. For example, you would want to analyse a customer’s sentiment scores (unstructured data) together with their buying patterns (structured data) when you predict their propensity to churn.

*When viewed as alternatives, the data lake and EDW models, are incomplete when it comes to addressing the kinds of challenges businesses currently face. The solution may lie with a hybrid of the two models.*

*Properly managed, a hybrid environment enables the implementation of a true contextual data lake, an evolutionary step up from the non-contextual data lake (the data swamp) to the real-time, virtual data lake environments to come.*

Having a data lake because it a cheaper way to store and manage data is not enough for business benefit - making the business money. A successful lake must be able to provide real-time response to queries and give users an easy and uniform access interface.

In this free data environment, enterprise data warehouses, discovery-oriented environments, and highly specialized analytic and operational technologies should coexist. By combining the best features of an enterprise data warehouse with the greater storage flexibility and contextual value provided by a data lake, businesses can more efficiently manage growing data volumes and complexity.

Executed properly, a hybrid environment can provide the ease of analysis you need today, while maintaining a repository of all data in its original format. Such an approach can address analytical needs while ensuring that context is preserved for both immediate and future needs.

An amalgamation of Traditional DW and Data Lake - a hybrid architecture - could be depicted as shown below.

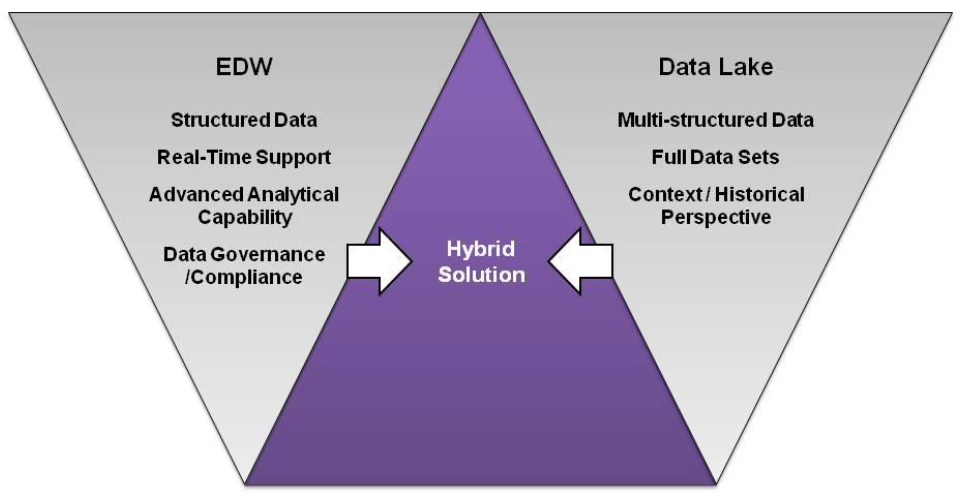
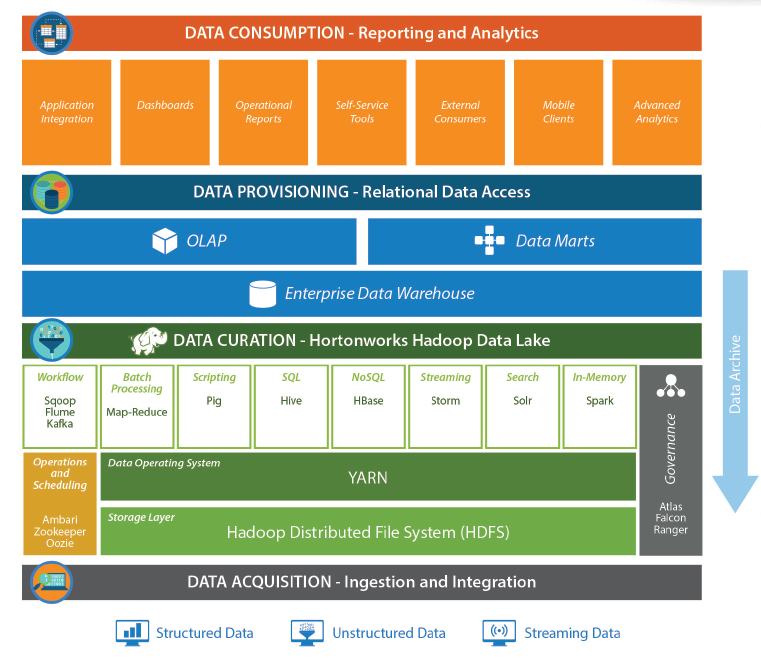


Figure 8: A hybrid solution to fill the gaps

A functional model of the hybrid data architecture is shown below. It describes layers within which data will be ingested, organized and presented to the business but it doesn’t specifically call out technologies that will be used to build these layers. This functional model aligns to physical layers within a final technical deployment.



#### Figure 6: Physical Layers of Modern Data Architecture – Example

Let us look at the different layers of the hybrid data warehouse shown in above sample architecture.

**Data Acquisition**

This layer refers to the ingestion and initial movement of data from the source systems whether they be traditional relational/transactional systems, user-generated data, unstructured or semi-structured data, external data or streaming data.

**Data Curation**

In the Modern Data Architecture, Apache Hadoop plays a key role as a data storage and curation layer. Using the data lake approach, all data – no matter what type – is stored in the data lake and is organized, shaped and made available for consumption by other layers.

**Data Provisioning**

Operational reporting and analytics are best served by more traditional data stores. The high-speed query capabilities of relational database systems make them ideal for serving data to support interactive query and analytics. Depending on the scale and needs of the organization, an Enterprise Data Warehouse built on a relational database platform may be coupled with several subject-oriented data marts to serve various reporting needs. In addition, an Online Analytic Processing (OLAP) engine can help facilitate complex, interactive query.

**Data Consumption**

This layer represents all end-user interfaces. It should be mentioned that although these physical layers may imply that there is no direct flow of data from the Curation layer to the Consumption layer, in some cases there is.

# Future of Data Lake

Increasingly, the world of big data is also the world of real-time data. Businesses no longer have the luxury of choosing whether they want to go big or they want to go fast. They must do both. This is where Data Lake can come into picture.

But, Data Lake should not be seen as a replacement for existing analytical platforms or infrastructure, instead, it can help the enterprise data architecture to evolve into an integrated solution. In this barrier-free ecosystem, enterprise data warehouses, discovery-oriented environments, and highly specialized analytic and operational technologies can coexist.

While the data lake can be viewed as an extremely valuable asset, many enterprise implementations succumb to the following five challenges:

1. No compelling business case
2. No compelling reason to move away from legacy tools and approaches
3. Lack of skills due to technology complexity
4. No appetite for change
5. Overall confusion and inability to get started

*Addressing the following five C’s can help enterprises fully embrace and gain business insight and value from the data lake and, ultimately, see a faster return on their investments.*

* **Case - Targeting ROI**

The first stumbling block is finding the right business case with a clear ROI. This is usually the number one reason why there is inertia around data lake adoption. Deciding who will pay for it and who will own, build, manage and maintain the data lake can easily become a challenge. Some of the most practical use cases for a data lake are data retention/archival/migration as part of a technology modernization initiative, new analytical applications, and a data hub for democratization of data within an enterprise.

* **Complacency**

Just like introducing any other architecture changes, the data lake faces its fair share of “don’t fix it if it’s not broken” or “not currently a priority” and even, “we don’t have any big data” syndromes. Same-old, same-old (“so-so”) data warehousing and legacy technology lacks the modern know-how to deal with today’s enterprise data and scale horizontally. Powered by the right technology, an enterprise data lake can be the key for the success of digital transformation initiatives and data-centric organizational models looking to create new revenue streams.

* **Complexity – With big data comes big skills**

Big data skills don’t come easy or cheap. Still a maturing market, emerging technology, skilled development and architecture professionals can be difficult to come by. Combine this with broad enterprise adoption of a data lake, working in conjunction with your current data management technologies and approaches, and it is much more complex than single, departmental project. Once past the infrastructure challenges, various moving parts of an ecosystem need to be pulled together that encompass software platforms and applications that don’t break each other.  One just needs to look at[The Hadoop Ecosystem Table](https://hadoopecosystemtable.github.io/) to get a sense of the complexity involved.

* **Change – The constant variable**

The large Hadoop ecosystem – with its active community – often keeps things interesting with constantly changing and evolving components. These changes are difficult enough for individuals to keep up with, but even more challenging for enterprises to adapt to – so much so that keeping up with the changes can be a full time job in itself. Without an appetite for innovation and a willingness to modernize your business processes, not just your technology, the promise of the enterprise data lake will remain elusive.

* **Confusion – Terminology and tools**

There are plethora of tools and solutions for the data lake which adds to the confusion around topics ranging from the exact definition of a data lake to learning best practices for the right technology for a use case, to data ingestion, management, governance, formats and even personnel roles. And we didn’t even mention the pending data tsunami that the Internet of Things (IoT) will represent in the future.

It’s important to remember that data lakes are well worth the effort and can produce tremendous value as a component of an enterprise’s data strategy. IT leaders that have the opportunity to kick-start a data lake initiative should prepare appropriately, getting “ahead” of skills gaps that will appear as the initiative evolves. Similarly, they should build their data management strategy with the cloud in mind, but remain flexible as IT environments become increasingly hybridized.

With these approaches in mind, enterprise IT leaders can successfully address the five Cs of big data management in the journey towards data lake adoption, bringing to life the profound impact and insights from business intel that will drive their businesses forward in the years to come.

Data scenarios suitable to consider Data Lake

Large social media companies like Facebook deal with hundreds of terabytes of information each day. A fleet of commercial jets can create similar amounts of data from a single day of operations. Traffic cameras, environmental sensors, and cell phones create and use untold billions of pieces of data that range from simple numerical data to voice, text, and video information.

Here are the ways that data lakes will influence the big data landscape in future:

* **Analytic Expansion:** By deploying data lakes, it is possible to place an organization’s data assets on an RDF graph, explaining the relationships between elements in such a way that effectively overcomes the “dark data” phenomenon, or data that goes unused by enterprises. Innately understanding the context and meaning of data prior to analysis affects the type, degree and nature of analytics performed, which considerably refines their results and use.
* **Semantics at Scale:** With semantics at scale, an organization utilizing a smart data lake graph is optimized for analytics with in-memory, massively parallel processing of semantically tagged data. Such an engine, when combined with a smart data lake’s RDF graph and ontological models of business meaning, incorporates all relevant enterprise data for comprehensive results at a speed which semantic technology advancements have only recently been able to produce.
* **Democratization of Stewardship:** The availability of data provided by data lakes is aligned with the self-service movement and the democratization of big data that supports it. Data lakes will contribute to the expansion of these trends by facilitating the democracy of data stewardship — a more pervasive form of governance than that conventionally reinforced by only a few dedicated data stewards. With increasing regulatory mandates, this enterprise-wide ubiquity of data stewardship will prove invaluable to organizations.
* **Automating IT and Data Science:** Additionally, the alignment of smart data lakes with the self-service movement will result in automation of some of the more mundane, but highly necessary aspects of data science and the work of IT departments. It will enable these professionals to concentrate on more substantial ways to improve data-driven processes.

# Conclusion

*Business related data is changing and increasing in volume and structure quite dramatically and hence calls for new approaches to handle the evolving data trends. Traditional data ware housing in current format will find it challenging to catch up to the velocity, veracity and volume of input data. A change is imminent and Traditional warehousing has to evolve to cater to the agile analytics demands.*

*That said, the data warehouse itself should stay as a logical representation of clean, vetted data that users at various levels can use to make decisions. Without a data warehouse, decision-makers operate by the seat of their pants, making critical decisions based on inaccurate or no data at all.*

*As the amount of structured and unstructured data increase, we may need to implement a data lake to complement the enterprise data warehouse. The contents and structure of the data lake will be determined by what data and analytics are required but which you cannot store or process in the conventional data warehouse architecture.*

*Considering the above, organizations can implements a BI ecosystem – a logical data warehouse in Gartner terminology – with a variety of technologies and tools, including a data lake. In this environment, the data lake and the enterprise data warehouse must both do what they do best and work together as components of the larger integrated logical data warehouse, which in turn, provides greater value by delivering insights previously not possible.*

*Hadoop, big data and the data lake does not replace a company’s existing investment in analytics. In fact, they complement it very nicely and will support, enhance modern analytics. Building a Modern Data Architecture that incorporates all of the benefits of a data lake, combined with the high-speed query and analytics provided by traditional relational data warehouse and online analytical processing (OLAP) engines, supports data consumption at all levels of the business. It also provides all classes of data consumers with the capabilities they require.*

*By building a Modern Data Architecture that suits the evolving data handling requirements, organizations can continue to leverage their existing investments in analytics, while collecting all of the data they have been ignoring or throwing away, all while enabling analysts to get company data and insights faster.*

*To conclude, bringing BI and analytics solutions together with the more robust discovery environment of data lakes can deliver more value to businesses. The combination of BI and analytics facilitates new business definitions and insights, and establishes new analytic pathways. Playing off the familiar Ralph Waldo Emerson adage that “Life is a journey, not a destination,” a data lake is both a journey and a destination because it is an architectural strategy and an architectural destination.*

# References

* The Technology of the Business Data Lake – *Author : Pivotal*
* How to design a successful data lake *– Author : Knowledgent*
* Data Lake Management Platform – *Author : Zaloni , the Data Lake company*
* Top Five differences between data lakes and data warehouses – *Author : Chris Campbell*
* Data Lake Vs Data Warehouse – *Author : Tamara*
* Data Lake Vs Traditional BI *- Author : Capgemini in association with Pivotal*
* Business Data Lake - *Author : Capgemini in association with Pivotal*
* How to Future-Proof Your Data Lake: Six Critical Considerations – *ITBusinessedge.com*
* 10 Things to Consider Before Diving Into the Hadoop Data Lake – *Craig Lukasik*
* Benefits of Hadoop based Data Lake – *Philip Russom*