Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

# Traditional BI vs. Business Data Lake – A comparison

# The need for new thinking around data storage and analysis

**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. The way forward is a complete rethink of the way we use BI - in terms of how the data is ingested, stored and analyzed.**

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.

**The answer is the Business Data Lake (BDL).**

A Business Data Lake is 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.

The Business Data Lake, particularly when combined with Big Data, enables business users to perform near real-time analysis on practically any data from any source. It does this by:

- Storing all data in a single environment (a cluster of data stores),
- Setting the stage to perform analysis (standard or self-service) on data whose structures and relationships are either already known or yet to be determined
- Providing analytical outputs for specific business needs across business functions
- Providing the capability to utilize the data for business benefits in near-real-time, with the ability to showcase data agility and enable agile BI.

# Traditional approaches and their pitfalls

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

## Top-Down Approach

The traditional Top-Down approach suggests bringing in the data from all the data sources, storing it in the EDW in an atomic format in a relational model and then building data marts with facts and dimensions on top of the EDW for analysis and reporting.

### Advantages

- Retains the authenticity and the sanctity of the data by keeping it close to the source form
- Provides a single view of the enterprise as all the data is present in the EDW.

### Disadvantages

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.

- A two-step process of loading and maintaining the EDW and the data marts
- Making data available from across the enterprise is difficult
- Time consuming as implementation is slow and the first output is usually several months away - businesses cannot wait that long to see the results.
- The business requirements may change by the time the implementation is complete.

## Bottom-Up Approach

This approach suggests bringing in the data from all the data sources, transforming and restructuring the data and loading them into the data marts in a dimensional model. The data marts will be subject-area-specific containing conformed dimensions across subject areas. The integration of the subject area specific data marts will lead to the EDW.

### Advantages

- Since this approach starts small and grows big, it is faster to implement than the top-down approach. This gives more comfort to the customer as the users are able to view the results more quickly.
- There is a one-step ETL effort of transforming/restructuring the data into the data marts, which may be high, but it brings down the 2-step process of data load of the top down approach.

### Disadvantages

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.

- The process of data restructuring (or transformation) while loading into the data marts involves complex ETL transformations.
- The process begins with small subject area specific data marts, which means gaining an enterprise level view takes longer. It does not, therefore, immediately address enterprise requirements.
- This approach is typically a collection of fragmented islands of information and may not really translate into an EDW.

The choice between implementing a top-down or bottom-up approach is usually based purely on the business need. In fact most organizations adopt a compromised, hybrid model which accommodates ideas from both approaches.

# An answer to the problem - the Business Data Lake

The Business Data Lake enables agile BI, thereby providing the capability to turn around the business outcome of the data consumed, in near-real time. Data agility enables provisioning of the right outcomes to the right users at the right time.

Big Data is all the rage. But what is it, ultimately? A collection of really large volumes / massive amounts of data that may be structured or unstructured, generated almost continuously and that is difficult to process or even handle using traditional data processing systems.

The Business Data Lake has been designed to solve these challenges around Big Data.

The data - both structured and unstructured - flows into the lake, and it stored there until it is needed - when it flows back out again.

Traditional BI systems leveraged the concept of a staging area. Here data from multiple data sources were Staged. This reduced the dependency on the source systems to pull the data. Data can be pulled at specific times into the Staging Area.

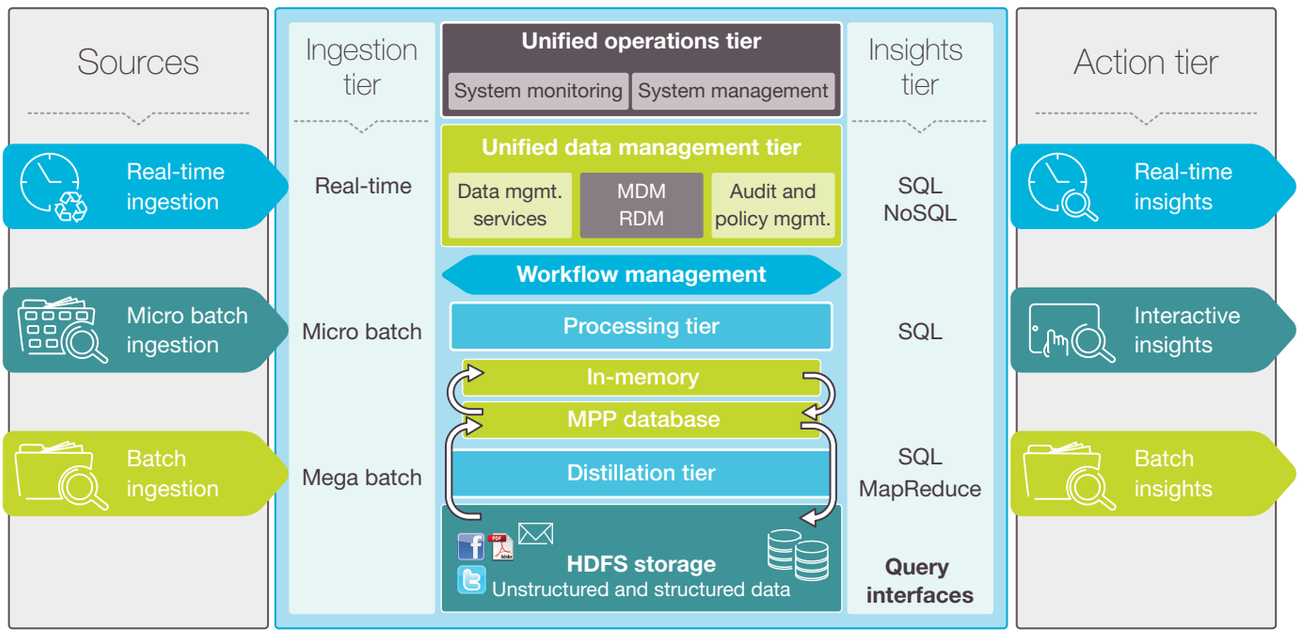
A Business Data Lake is very similar to the staging area, the key difference being that the Lake stores both structured and unstructured data. It also provides the ability to analyze data as required by the business. Any kind of data from any source can be loaded into the Lake. There is no need to define structures or relationships around the data that is staged. In addition, storage appliances like Hadoop can bring in all the enterprise data, without worrying about disk space or judging whether a piece of data is required or not.

In a traditional DW implementation, the data in the staging area is transient. There is no persistence of data. It has not been possible to stage such large volumes of data for extended periods of time due to hardware costs and storage limitations. In the Lake, the limitation on the storage is eliminated by using commodity hardware that is much cheaper. Thus, the data that is staged is persistent over time and non-volatile.

Traditional DW approaches require long processes of data ingestion. It can take months to even review results from the data. The Business Data Lake enables agile BI, thereby providing the capability to turn around the business outcome of the data consumed, in near-real time. Data agility enables provisioning of the right outcomes to the right users at the right time.

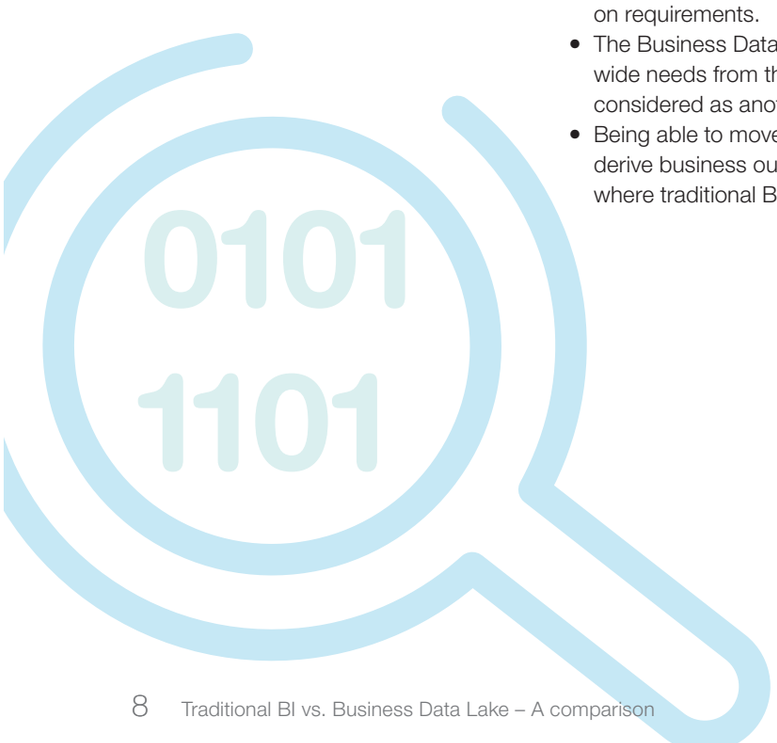
Governance can be implemented on the data residing in the Business Data Lake as required. Master data definitions and management can happen on the data that is required for analysis, thereby eliminating an over-engineered metadata layer, providing for data refinement/enrichment only for relevant data.

Business Data Lake Architectre

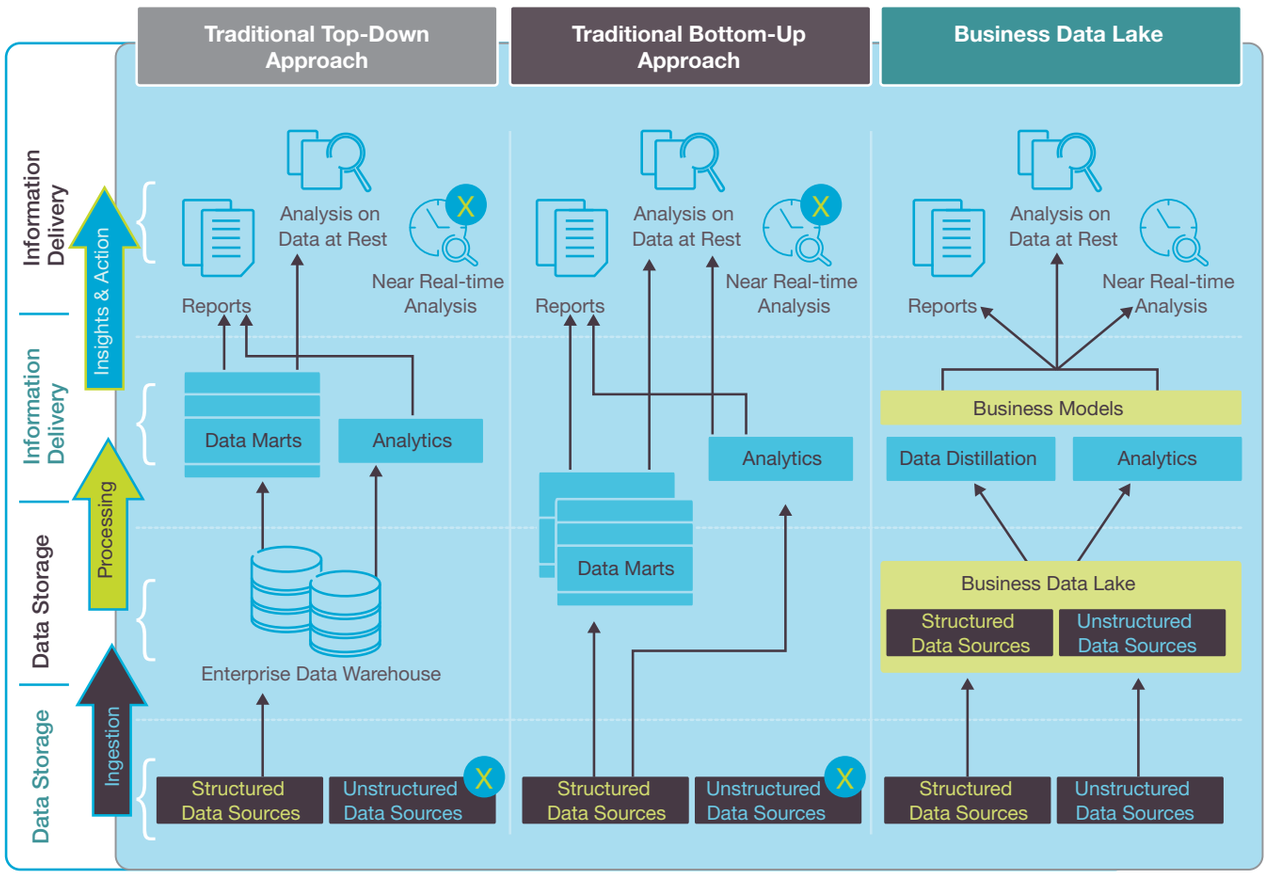



# Benefits of the Business Data Lake

- 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.



# Architecture Comparison — Traditional BI and Business Data Lake



| Tier | | Business Data Lake | | Top-Down (EDW) | Bottom-up (Data Mart) |
|---|---|---|---|---|---|
| Storage | Process | ALL Data | | Structured data | Structured data |
| | Cost | Low | | High | Medium |
| | Effort | Low | | High | Low |
| Ingestion | Process | ALL Data Sources | | Multiple structured Data Sources | Multiple structured Data Sources |
| | Cost | Low | | High, due to effort | Medium, due to effort |
| | Effort | Low, as all data flows in to the lake | | High, due to data alignment into EDW | Medium, due to data alignment into data mart |
| Distilation | | Done on demand based on business needs, allowing for identifying new patterns and relationships in existing data.<br>**This process is a differentiator** | | Already distilled and structured data, does not allow for further distillation | Already distilled, structured and aggregated data, does not allow for further distillation |
| Processing | | Capable of handling analytics on the data in the lake<br>**This process is a differentiator** | | Not possible directly on the EDW | Not possible directly on the data mart |
| Insights | | Ability to analyze data as required. Allows for data exploration and so enables the discovery of new insights that were not directly visible | | Analysis needs to be defined upfront and hence is rigid to the business need | Analysis needs to be defined upfront and hence is rigid to the business need |
| Action | | Ability to integrate with Business Decisioning systems for the next best action | | Technically feasible, but not effective due to data latency | Technically feasible, but not effective due to data latency |
| Unified Data Management | | MDM and RDM on relevant data. | | Effective MDM and RDM strategies exist, but possibility of over-engineering | Effective MDM and RDM strategies exist, but possibility of over-engineering |

COMPARE TO

As we see, a Business Data Lake is able to:

- Receive and store high volume and volatile structured, semi-structured and unstructured data in near-real time using low cost commodity hardware
- Provide a platform to perform near-real time analytics and business processing on the data in the lake
- Provide a business view that is tailored to specific LOBs as well the enterprise.

The Business Data Lake does this in a way which enables users to reduce the business solution implementation time, by:

- Eliminating the dependency of data modelling up-front and thereby letting all data flow in
- Reducing the time taken to build robust ETL process to load the data into the structured data stores, which are bound to change
- Eliminating an over-engineered metadata layer
- Providing the capability to view the same data in different dimensions and derive new patterns and relationships that lie within the data.

**A Business Data Lake is a simple but powerful approach to solve business problems. It caters to ever-changing business needs by allowing for storage of all data and providing the capability of deriving actionable insights from any kind of data, yet working in a transparent and seamless fashion, in an enterprise-wide environment.**

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™


## About Capgemini

With almost 140,000 people in over 40 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2013 global revenues of EUR 10.1 billion.

Together with its clients, Capgemini creates and delivers business and technology solutions that fit their needs and drive the results they want. A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

Find out more at **www.capgemini.com/bdl**
and **www.gopivotal.com/businessdatalake**

Or
contact us at **bim@capgemini.com**

MCOS_GI_AP_20140805