Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

The Principles of the Business Data Lake

The Business Data Lake

'Culture eats Strategy for Breakfast,' so said Peter Drucker, elegantly making the point that the hardest thing to change in any organization is its culture. In the areas of information management, analytics and reporting, however, IT has been continually fighting a business culture which it has no ability or authority to change. This has been done in part due to two physical constraints: the cost of data storage and movement, and one IT imposed constraint: the Single Canonical Form.

The impact of these constraints has been that while IT has pushed towards single central EDW (Enterprise Data Warehouse) solutions, the business has continued, even increased, the use of Excel spreadsheets and isolated solutions. The challenge for IT is that as the value of information has increased so the motivation of the business to accept slowly evolving single solutions has decreased. The Business Data Lake looks to solve this challenge by using new Big Data technologies to remove the cost constraints of data storage and movement and build on the business culture of local solutions. It does this within a single environment – the Business Data Lake.

The Business Data Lake is not simply a technology move. It is about changing the culture of IT to better match the business culture. The historical battle between business unit independence and the centralizing ambitions of IT and corporate management has proven to be an unwinnable war.

The Business Data Lake addresses the challenge by building a single culture and concentrating on the areas that deliver true value.

The Business Culture of Information

Businesses have always had a thirst for information. That thirst has evolved from being merely important to the operations of the business through to its position today as the foundation of strategy and success. The challenge for IT has never been the desire of the business to access and view information – instead it has been the continual battle between IT and the business on how that should be achieved.

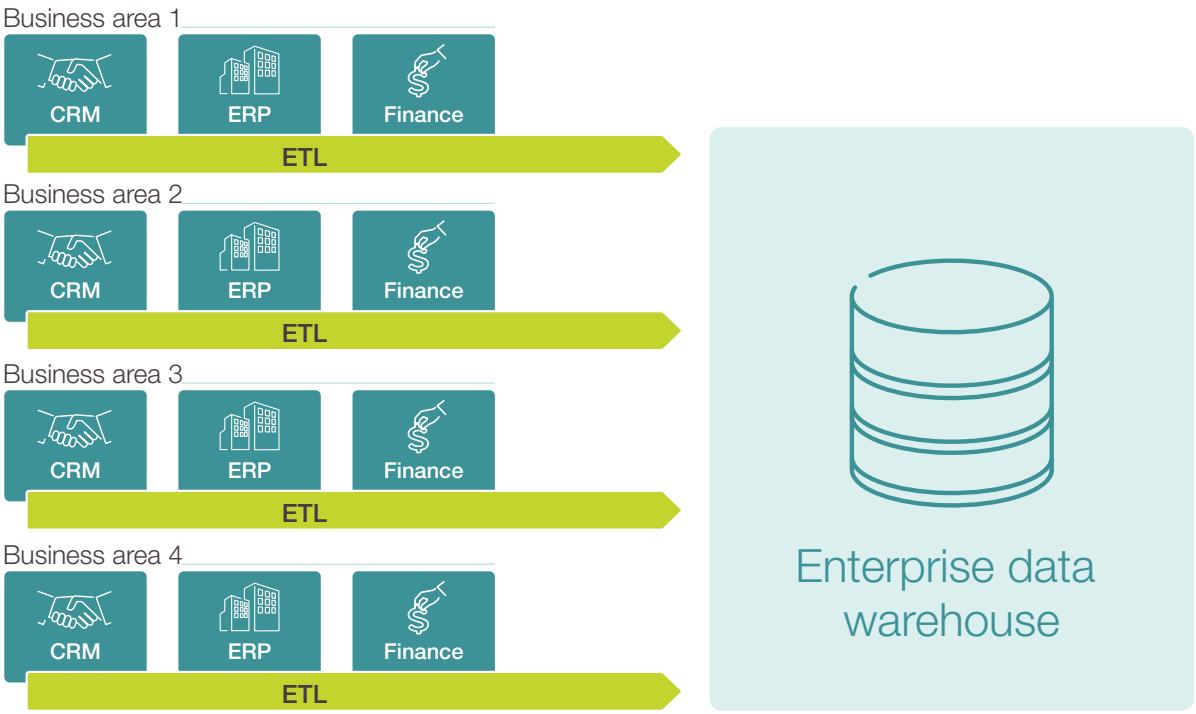
On one side a business is organized as a combination of departments, regions or verticals, each with its own management structures and KPIs. At the Corporate level there is the desire to obtain a consistent view of the *corporate* KPIs. The business culture of information is therefore about *local* views of information, *local* perspectives displayed in a *local* context. Even at the corporate level the desire is to see views localized to the corporate level. This business culture encourages the use of point solutions and often discourages more horizontal governance. If something is 'right' for an area there is less motivation to enable the horizontal view.

The IT Culture of Information

Into this Business culture IT has historically tried to establish a *'one size fits all'* solution – the Enterprise Data Warehouse. The goal of this is beguilingly simple: *Create a single place with all of the enterprise information so governance, modeling and loading only needs to be done once.* Unfortunately this vision has one key constraint: *Everyone in the business needs to agree what they want to see.*

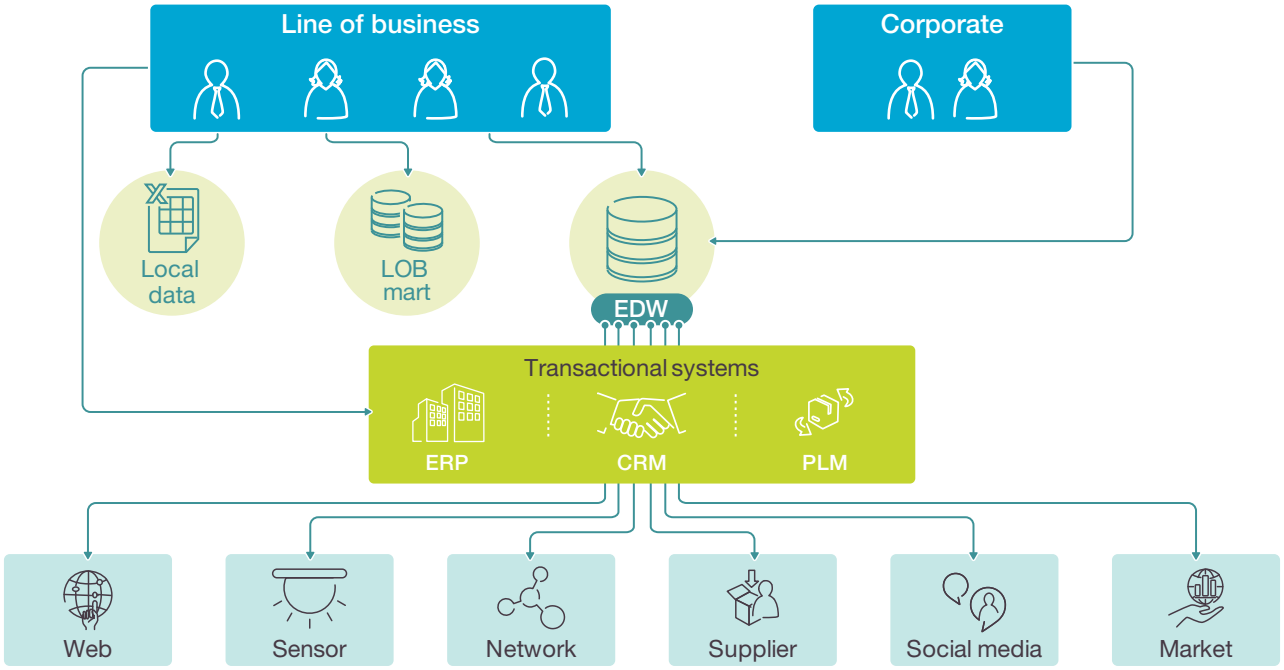
Figure 1: Multiple corporate views

Figure 2: Centralization view



This centralization view is intended to give the business whatever it needs and provide a single point where all business users can go. This IT culture drives centralization and attempts to stop Line of Business (LOB) data marts and reporting solutions. The reality however is that very few, if any, businesses have actually removed all Excel use and LOB marts. Instead, these point solutions are often the most used information applications.

Figure 3: Data consumed through multiple internal views



This culture can be summed up in a few key principles:

1. There is a single view for everyone and everything – The Single Canonical Form
2. Move only what is required
3. Push the business to the single view
4. Prevent point information solutions

The data warehouse methodology hasn't changed for 30 years

Most data warehouses have been designed using a central layer based on a 3rd normal form modeling (often named single canonical form). This layer is supposed to be fed by every source (both internal and external) and to be used to feed any data mart or usage.

It is this layer that enforces the "single version of the truth" way of working. Any information put in the data warehouse is heavily and precisely defined, follows group processes and has been approved by both business, IT and data officers, and often corresponds to a consensus, which means nobody gets the view they want but everyone is equally unhappy.

This way of working can work within siloed group functions, such as Finance. But three main issues appear when attempting this more broadly:

1. The time taken to add new information or to update an existing business process or KPI is very long (from weeks to months) and much slower than the business needs
2. Aligning local Business Units and group processes is very complex, very long, and often not possible or desirable
3. When dealing with external information the ability to enforce this standardization is nearly, or completely, reduced

The result of this compromise is then industrialized and fixed by IT through the ETL (Extract, Transform, Load) process. Here IT compounds the problem by having a simple philosophy, one that is done with the best of motives: *reduce cost.* The goal at this stage is to just extract the *minimum* of information that matches the schema. The ETL process extracts this minimum and then applies processes which endeavour to enforce consistency across all areas, thus ensuring that the unhappy concensus is fixed at all stages of the process.

The driving imperative behind this approach has been the reduction of cost, by creating the single view and then concentrating on just that view it becomes possible for IT to minimize the *IT* costs of the solution. However the impact of this unhappy consensus and IT approach is simple: Business Units build their own decision platforms (often with their own Data Warehouse) and commit only to feeding the minimum information to the corporate data warehouse but not use it.

Today's challenges aren't the challenges of 30 years ago

The old approach was based on the challenges of 30 years ago, multiple lifetimes in an IT sense. Today there are many more questions around data that need to be answered:

- How to handle unstructured data?
- How to link internal and external data?
- How to adapt at the speed of business change?
- How to remove the repetitive ETL cycle?
- How to support different levels of data quality and governance based on differing business demands?
- How to let local business units take the initiative?
- How to ensure the platform will deliver and will be adopted?

Combined with this the last 30 years has seen a dramatic change in the technology available. Technologies that can slash the cost of data storage, enable real-time analytics and provision information for business users at much faster speeds. It is these new challenges and the impact of new technology that has led to the Business Data Lake solution and methodology. An approach that starts with the objective of building on how the business operates and delivering a new information culture that leverages rather than fights the business culture. The Business Data Lake is built for today, using today's technology in a way that meets the demands of business today.

The Principles behind the Business Data Lake

The Business Data Lake changes the way IT looks at information in a traditional EDW approach. It embraces the following new principles:

1. Land all the information you can *as is with no modification*
2. Encourage LOB to create point solutions
3. Let LOB decide on the cost/performance for their problem
4. Concentrate governance on the critical points only
5. Consider the corporate view to be just another LOB view
6. Unstructured information is still information
7. Never assume the lake contains everything
8. Scale is driven by demands – scale down as well as up

These new principles drive a new approach, one that delivers what IT needs – *a cost effective solution* in a way that leverages the business need for local views.

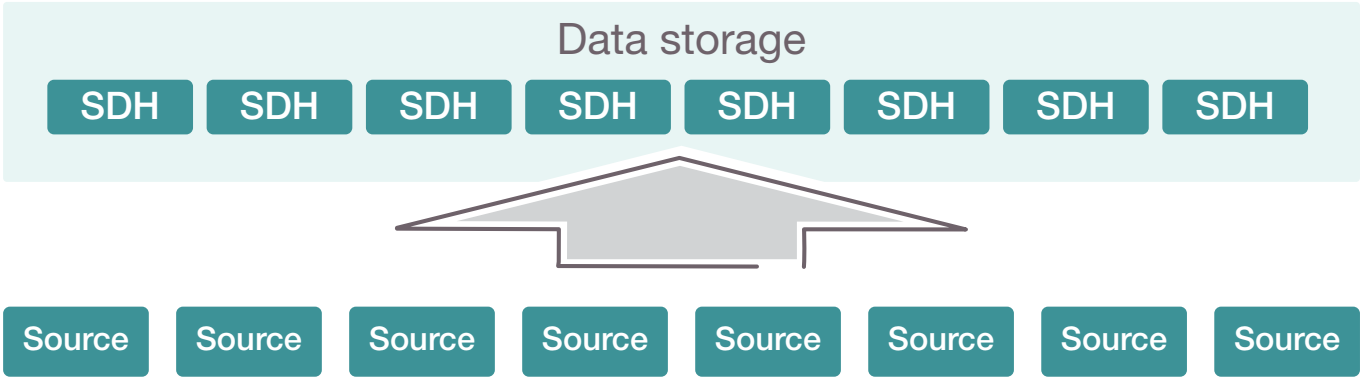
Land Everything – Data Storage

The first change with the Business Data Lake is the desire to land everything *without modification*. Using the Hadoop file system (HDFS) it is possible to simply 'dump' information from source systems into Hadoop and not worry about transformations or formatting. This new approach means that:

- Time analysis of information is now possible
- Information maps can be left until needed, and are no longer required at the start of a program before data can be ingested.

This new approach is extremely quick to deliver as the technical complexity is low. It also means that IT *has already* made the information available for the business to use, and more than just the current information it contains the full source data history (SDH) of those source systems.

Figure 4: Data storage



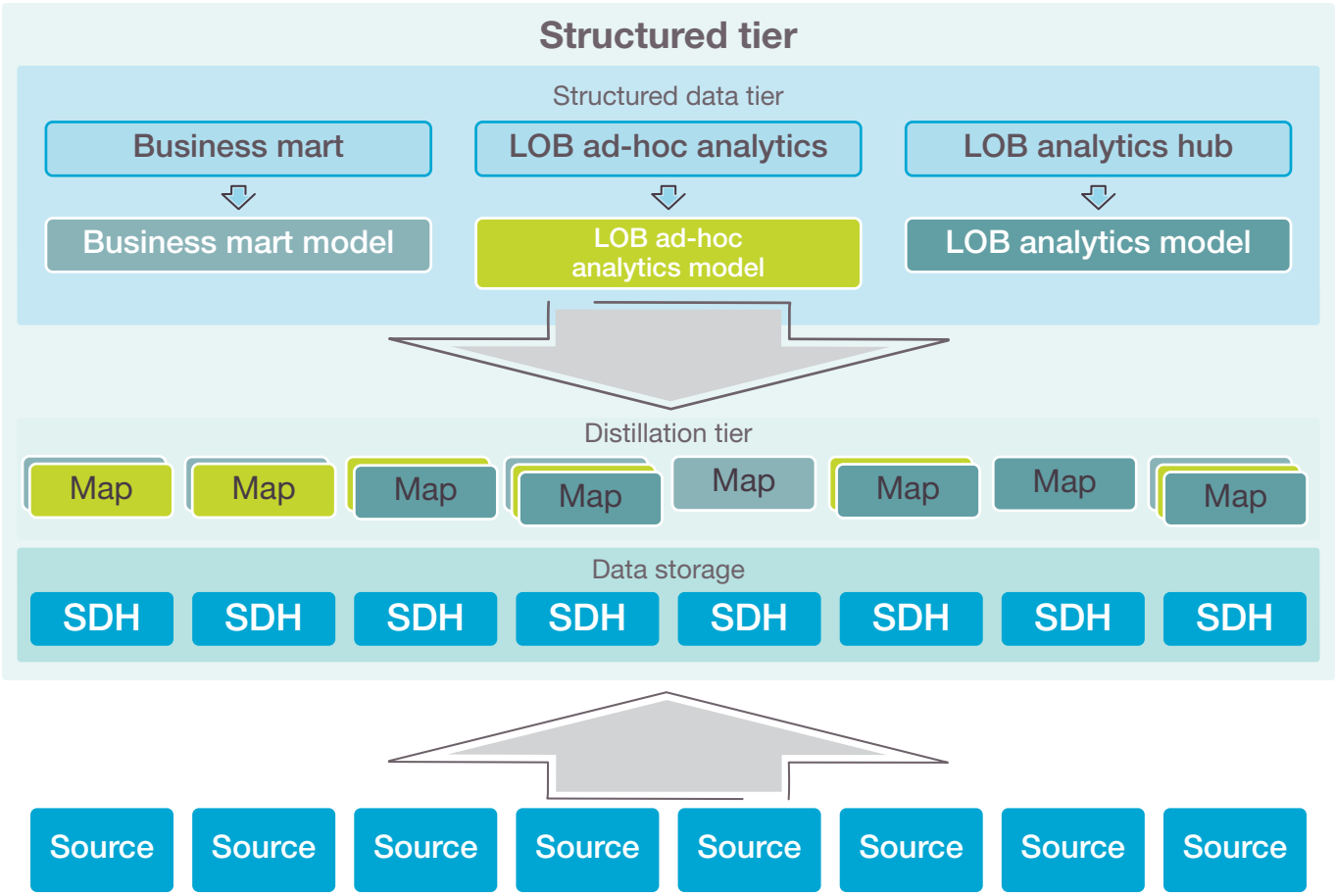
Encourage LOB and Let them Decide – Distillation

The next key part of the Business Data Lake is the concept of *distillation*. This is where the business creates maps onto the source data histories contained in the Data Storage tier to generate the view that matches their current requirements.

The goal here is to enable the business to extract any information they are allowed to: privacy and security can be enforced through the distillation process. These maps can be reused by others or just discarded, as can the point information solutions if required.

By providing the business with access to all of the raw information, operational reporting systems can now be created in the same environment as long-term financial planning and corporate reporting. Critically, this removes the business need to create point solutions: if all the information is already there, why bother creating your own?

Figure 5: Distillation tier



Choosing Performance and Cost

Currently in technology there are two choices for structured analytics:

- Disk-based structured analytics which can handle large volumes, performs well and is cheap but relatively slow
- In-memory analytics which is incredibly fast but more expensive

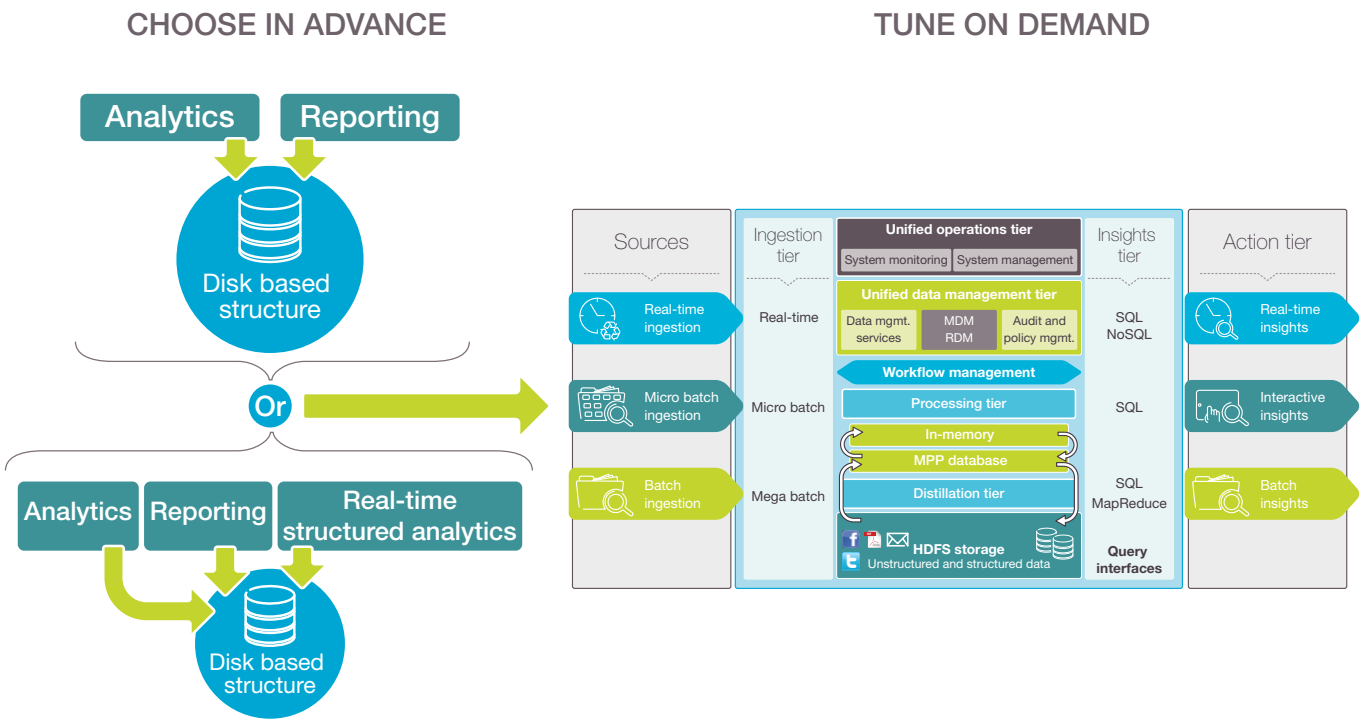
Most annoyingly for business users, these two approaches are normally in completely different technology stacks, so the decision on 'cheap v fast' is one that is made before the business really knows the return. This results either in a solution that is too slow, or one that is too expensive.

Governance where it counts - Collaboration

The Business Data Lake does not mean that information governance is not required. Instead it means there is a greater business reason for governance and that the governance can be concentrated where it delivers value: *governance counts when people need to collaborate.*

Simply put, this means that the Business Data Lake provides a platform for the collaboration but the business needs to agree on *what is required for collaboration*. In most organizations this is a small amount of cross-business information – the Master and Reference Data – and minimal amounts of transactional data. This approach, termed the *Minimal Canonical Form* aims to define the very smallest set of data which enables different business views to collaborate.

Figure 6: Choosing performance and cost



The Business Data Lake does away with this problem by containing *both* disk-based and in-memory approaches in a single infrastructure. Thus the business can prototype and prove value on the cheaper infrastructure before switching to in-memory or even combine both approaches in a single solution. The Business Data Lake is about enabling 'elastic analytics' for the business, linking cost and performance decisions to the business value delivered.

As an example, if the Warehouse team needs to understand the sales forecast by product in order to better manage stock then they need to agree on:

- How to uniquely identify products in a consistent way (Master Data)
- How to measure quantities (Reference Data)
- How to define regions and dates (Master Data)

They simply need to see the sales forecast at this coarse level, not to understand all of the details of customer names, sales people, contacts, marketing campaigns and other

elements. Collaboration therefore is about being *precise in communication only where it adds business value.*

This new approach to business information governance builds on the business culture and provides IT with a platform through which to industrialize it. By including Master and Reference Data and the cross references between business areas in the Business Data Lake, it becomes possible for business users to create ad-hoc collaborations and to construct views that truly represent what each area wishes to know about another.

The corporate view is just another Line of Business

Traditional EDW Single Canonical Form approaches have taken the perspective that the corporate view is one that includes everything. The reality is that the corporate view wants to see the *aggregation* of the LOB view, but the only historical way of achieving this would be to put all of the detail into the corporate view to enable that aggregation.

Because the Business Data Lake enables LOB solutions and focuses governance on collaboration, a new approach becomes possible: to set corporate standards for data reporting in the same way as financial reporting standards. Each business unit will already have a set of financial reporting standards it must meet and the Business Data Lake gives the opportunity to expand that approach and define what is the *minimum* set of information each LOB must include within their views to enable the corporate view to be an aggregation of the LOB views.

This approach leverages the corporate culture of financial reporting. It also enables the corporate level to improve the way it sets, disseminates and reports on KPIs. By concentrating on this minimal approach, the corporate level can focus on what is most important.

Unstructured Information is Information too

The Business Data Lake is at its heart a full platform for information, which means although it excels at all types of structured information, it also treats unstructured information as a first-class citizen. The Business Data Lake enables both the performance of unstructured analytics and the mapping of unstructured information into structured data.

By handling unstructured information within the same platform, and providing a low cost mechanism for accessing and processing it, the Business Data Lake makes it possible for business users to take advantage of new, unstructured Big Data analytics and to combine that with their normal structured data reporting. With the Business Data Lake there is no reason to shift between environments or for the business to create point solutions. In this way IT provides the answer to the business question 'how will we deal with large volumes of unstructured data?' The answer is: 'within the same environment and leveraging the same tools.' Historical approaches have had to separate unstructured analytics from structured reporting. With the explosion of Big Data this approach is no-longer viable and it is particularly in the unstructured data area that traditional approaches completely break down and local point solutions are the only option.

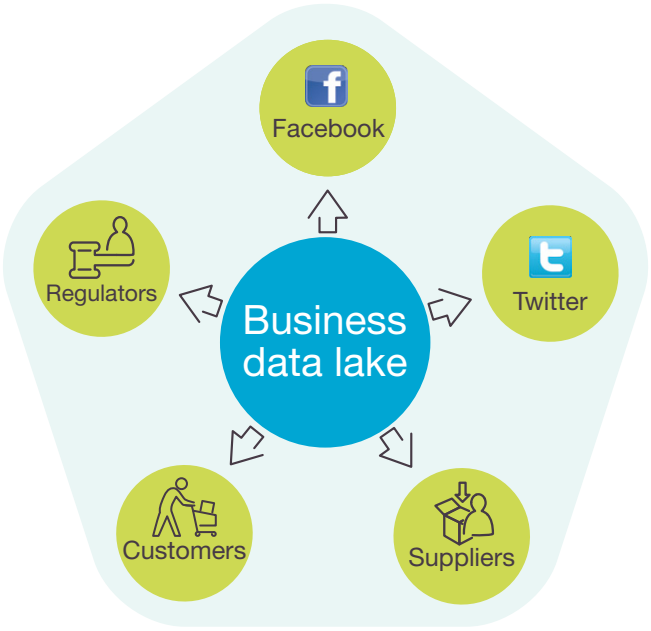
Figure 7: The corporate view

The Business Data Lake is designed from the ground up as a Big Data solution, and for that reason *all information* is managed in a single environment.

Never assume you have everything

The final key principle of the Business Data Lake is that you should never assume you have everything. If you are looking at Social Media data you are not going to be able to copy all of Facebook and Twitter into your Business Data Lake. Therefore the ability to *federate analytics* and combine the results within the data lake becomes important. For this reason the Business Data Lake contains tools to move and manage external information as well as federate analytics and collate results.

Figure 8: Federated analytics



Scale is driven by demand

The other big shift from a traditional EDW solution is the ability of the Business Data Lake to run on commodity or cloud infrastructures. This is done for two reasons: firstly to reduce costs and secondly to better link scale to demand. Rather than scale being driven by large appliance or large server purchases it can be done incrementally and in the case of cloud solutions can include the ability to scale down as well as up based on demand. This sort of flexible infrastructure helps to better manage IT costs and deliver to the business a more cost effective 'elastic analytics' capacity where decisions on scale are driven by current rather than assumed demand.

Business Culture driven Information Management

By examining how business leaders use and interact with information and leveraging next generation information technologies it becomes possible to change the constant battle between IT's centralized view and the need for business agility and localization. By providing a platform which builds on business culture and focusing governance where *justified by business value* it becomes possible for IT to demonstrate how it adds value to the business and do so in a way that manages *cost* in a centralized manner while also providing the business with the flexibility that it needs.

The Business Data Lake is a new approach to information management, analytics and reporting that better matches the culture of business and better enables organizations to truly leverage the value of their information.

With the value of this information only set to increase, businesses cannot afford to continue with approaches that have been proven to fail and to be less than agile even in the small areas where they have been forced to work. The Business Data Lake is a new approach enabled by a new generation of technologies. No longer do organizations need to be weighed down by the technology constraints of the past. But more than a technical solution the Business Data Lake is a new approach to the information challenge based on the principle that collaboration and access to information lie at the heart of effective business.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

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

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

About Capgemini

With more than 130,000 people in 44 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2012 global revenues of EUR 10.3 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.

MCOS_GI_MK_20131126