WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

Trends in Establishing a Data-Driven Enterprise

By Mike Ferguson
Intelligent Business Strategies
May 2020

Prepared for:

denodo

Table of Contents

WHAT IS A DATA-DRIVEN ENTERPRISE?

It is common in many organizations today to hear that they are looking to become a “data-driven” enterprise. But what exactly does that mean?

A data-driven enterprise maximizes the use of data and analytics to drive business value

A data-driven enterprise maximizes the use of data and analytics to guide decision making to achieve the maximum possible business value, such as improving the customer experience, optimizing business operations, or reducing fraud. It seems that while a lot of the spotlight in a data-driven enterprise is given over to the benefits of machine learning and artificial intelligence, the insights produced are useless if the data they rely on is poor.

The quality of data is critically important if decisions dependent on it are to be effective

We have all heard the expression “garbage in, garbage out.” Therefore, the data needed to drive the most optimal decisions has to be of the highest quality and be complete in the sense that all the data needed from one or more sources needs to be available to maximize the desired outcomes.

Companies need to organize themselves to become data-driven in order to produce data products for others to use in driving value

In that sense, a data-driven organization is one that organizes itself to become data-driven. It needs to be set up to produce and publish trusted, re-usable data products that others in the organization can easily find, access, and use to produce the necessary analytics and business intelligence needed to drive decisions to achieve those outcomes.

However, there are many challenges in the way of achieving this, all of which need to be overcome if companies are to meet the business expectations of a data-driven enterprise. It follows therefore that there are also a number of critical success factors to help make this possible. This paper looks at these, as well as the key trends in companies striving to achieve this status and then looks at the vital role that data virtualization plays in enabling companies to become data-driven.

What Are the Challenges to Becoming Data-Driven?

In this era of digital transformation, companies need to overcome a number of challenges if they want to engage in data-driven decision making.

Operating Environment

The first and most obvious of these is that the operating environment that many are now using to run their business includes the data center, one or more cloud computing environments and also edge computing made up of thousands of devices. This is shown in Figure 1.

We are now dealing with a hybrid operating environment that includes data centers, multiple cloud computing environments and edge computing

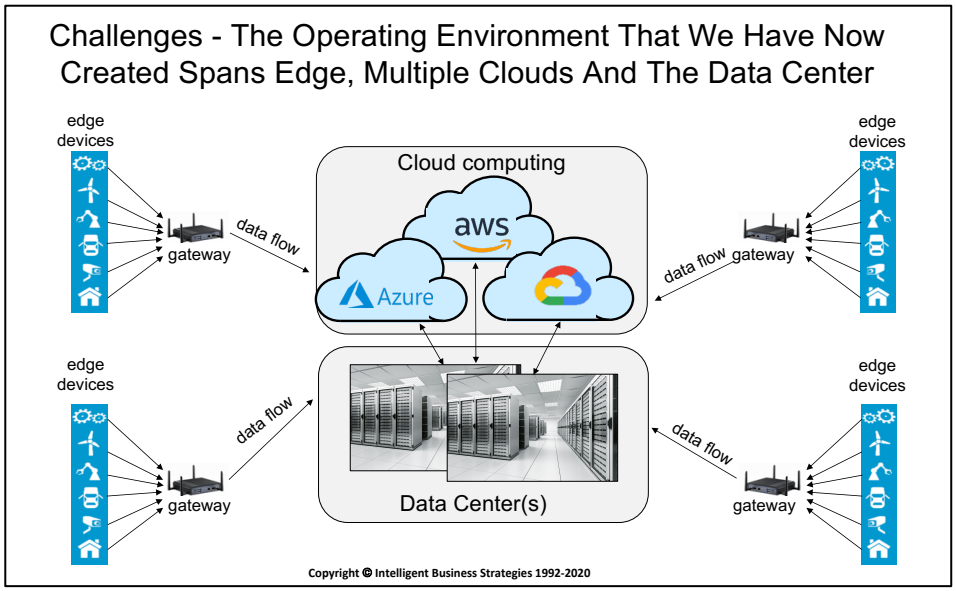

Figure 1

Data Complexity

Data is incleasingly becoming more and more distributed across multiple different kinds of data stores in multiple locations

Not surprisingly, in this kind of environment, data is captured and persisted in a wide range of data stores including relational databases, NoSQL databases, streaming message queueing systems, file systems, cloud storage, Hadoop systems, and edge databases. It is a distributed data landscape.

A distributed data landscape increases the risk of data duplication, inconsistency, poor data governance, and overspend on data tools

This kind of environment opens up a number of issues. For example, data resides in multiple data stores across this landscape. It is highly likely that there is data duplication, overlapping subsets, and multiple versions of data in existence. In many cases, people may have no idea where data is or what it means. Also, people may have no idea of data quality, how access security is governed, or what data is processed where and by whom. An obvious question is therefore, how do you capture, store, integrate, analyze, and govern data across edge devices, multiple cloud computing environments, and one or more data centers? What happens if it is too big to move from where it is stored or cannot be moved for legal reasons? How do you integrate it then? Also, how do you stop people repeatedly integrating the same data potentially inconsistently with many different tools?

WHAT ARE THE CRITICAL SUCCESS FACTORS?

These are just a few of the issues and questions that can arise in a modern enterprise. They highlight real obstacles to becoming data-driven. Given the backdrop of an increasingly complex distributed data landscape spread across data stores in the data center, multiple cloud computing environments, and the edge, several critical success factors need to be in place if organizations are to avoid the chaos stemming from different departments buying their own self-service data preparation tools and creating a "wild west" situation like that shown in Figure 2. We cannot afford everyone blindly integrating data with no attempt to share what they create and potentially re-inventing the same or similar data sets inconsistently over and over again.

Ungoverned puchasing and use of self-service data preparation across departments can lead to data chaos

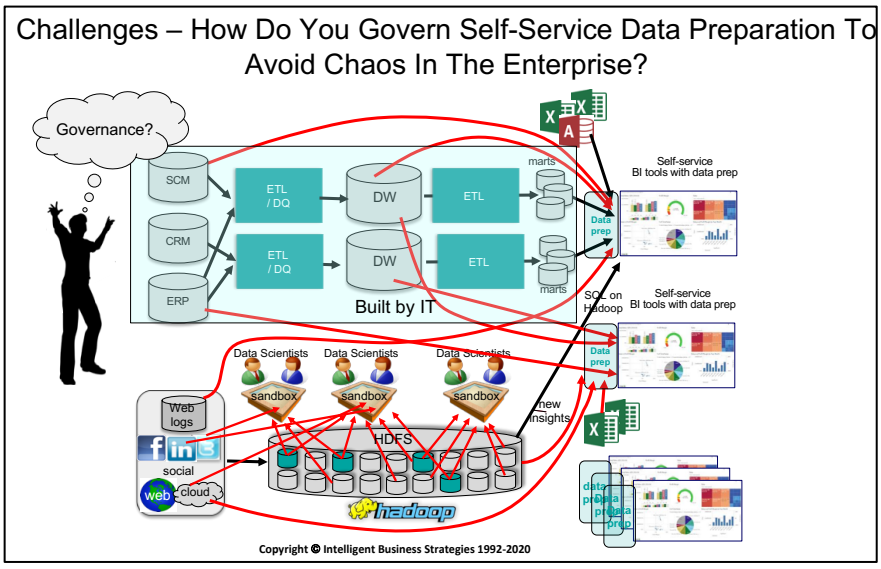

Figure 2

If companies are to overcome the challenge of complexity to become a data-driven enterprise then the following critical success factors need to be in place:

Establish a data culture and a data governance framework, and buy data tools that work together

A data lake, data catalog, and skilled data professionals who can engineer data are very important

Critical Success Factor	Reason
Data-driven culture, organized program office, approved projects, and project teams	Mass participation in a common approach to becoming data-driven
Data governance framework	Govern data quality, privacy, access security and retention across a distributed landscape
Integrated data tools	Productivity - Data catalog, data fabric (preferably including data virtualization) that can connect to and access data across a distributed data landscape that encompasses the data center, multiple clouds, and edge devices
Data supply chain - data lake to data marketplace with virtual data provisioning	Consistent approach to building and provisioning data and analytical assets
Data skills	Trained business and IT professionals that can hit the ground running

Component based development using orchestration and common data services is needed

Common data names and definitions	Consistency across data models, XML/JSON schemas, data virtualization virtual views etc.
Common transformations	Get consistency in data integration
Common data quality rules	To cleanse the same data in a consistent standard way no matter how that data enters or flows across the enterprise
Common data integration services	To transform and integrate data in a consistent standard way
Common governance policies	Data quality, privacy, access security, retention
Common trusted data services	To provide trusted consistent data to whoever or whatever needs it *preferably virtually* to avoid creating multiple copies of the same data
Analytical models	Shared across one or more applications

Data sharing and reuse improves productivity and shortens time-to-value

These critical success factors are to enable data governance while maximizing the ability to share and re-use data to improve productivity and shorten time-to-value.

WHAT ARE COMPANIES DOING TO BECOME DATA DRIVEN?

Given these critical success factors, what are companies doing to become data-driven? The following sections show some trends from the 4th Industrial Revolution Survey conducted over 500 enterprises in November 2019.

HOW ARE THEY ORGANIZING TO SUCCEED?

CIOs and CDOs are the most popular executives accountable for data

There is no question that one of the hardest problems to solve in establishing a data-driven enterprise is how you organize to succeed. 19.6% of companies surveyed stated that the CIO/CTO was the most senior executive responsible for leading their data-driven initiative. The Chief Data Officer (CDO) was the second most popular senior executive responsible. However, in large enterprises over $1bn in size, the CEO is leading the data-driven initiative with the entire board and corporate legal department now also involved.

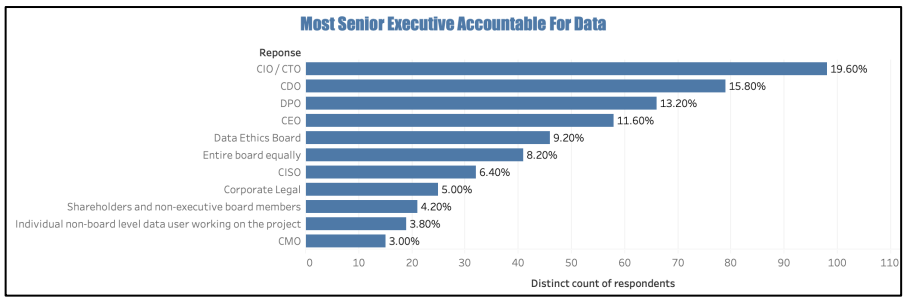

Figure 3

In addition, a federated organization structure is being used to organize project teams across the business to produce data and analytical assets to help achieve specific strategic business objectives using common technologies.

A federated organization is being established to organize projects across the business and align them with strategic business goals

Project teams are being encouraged to use common technologies for data engineering and analytics

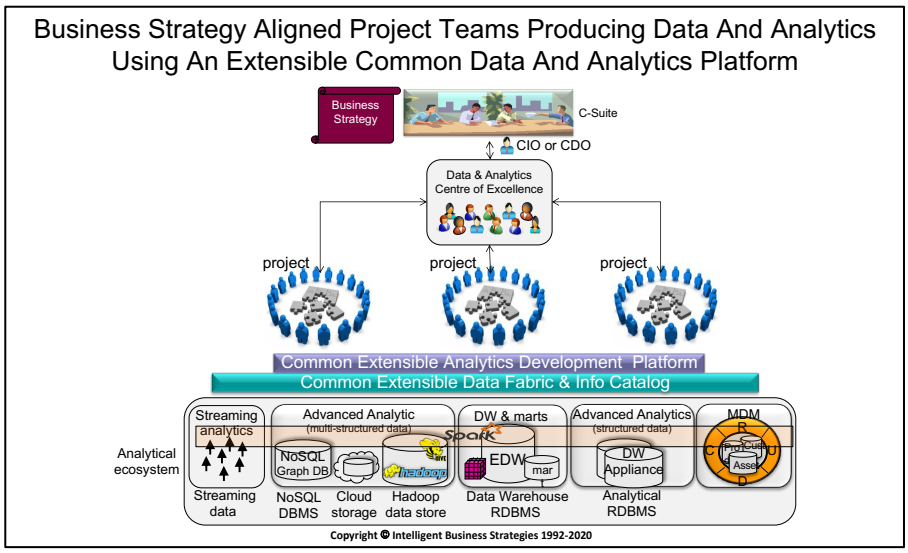

Figure 4

HOW LONG IS IT TAKING TO ESTABLISH A DATA DRIVEN CULTURE?

Peter Drucker famously said “Culture eats strategy for breakfast” meaning that culture can kill off any strategy within a business. This is because it really requires employee buy-in for a data-driven culture to take hold.

Figure 5 shows how long it is taking companies to get most of their employees to see the value of a data-driven approach. 34% of companies say that it is taking between three to six months while 30.8% indicate six months to a year.

It can take 3-6 months or even up to a year to establish a data-driven culture

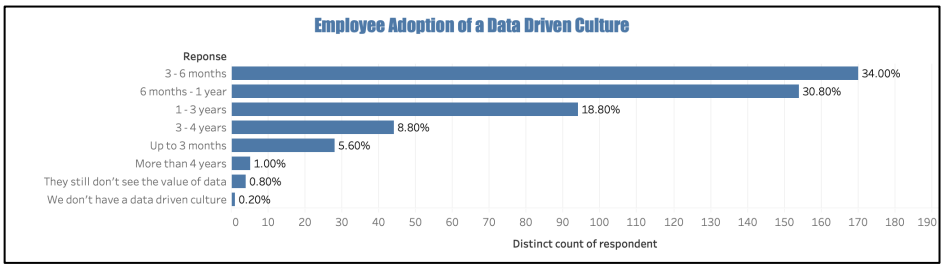

Figure 5

WHAT CHALLENGES AND INHIBITORS ARE COMPANIES FACING IN TRYING TO CREATE A DATA DRIVEN ENTERPRISE?

A number of key challenges in creating a data-driven enterprise were cited by companies. With respect to implementing a data culture, the top two challenges were the budget needed to upskill current employees (46.8%) and gaining understanding as to why this is important across the whole organization (42.8%).

Getting the funding to upskill and getting buy-in across the whole organization are key challenges

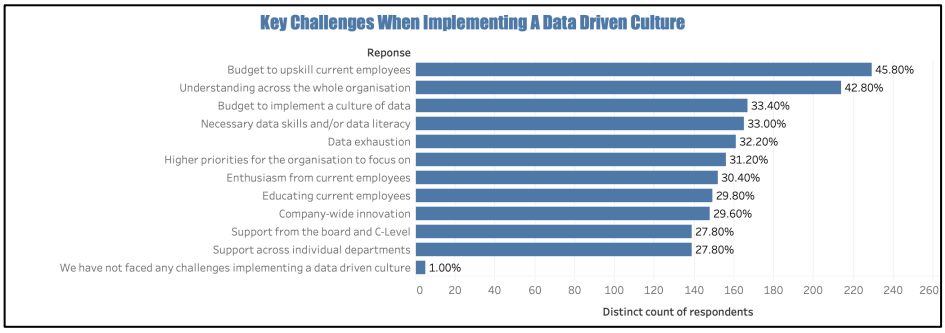

Figure 6

There are also skills issues. 43.4% of companies surveyed said that lack of technical skills in cloud computing in areas such as strategy, cloud migration, cloud deployment etc., were preventing them achieving their business goals. In addition, 43% said that a lack of data engineering technical skills was also a major problem, as shown in Figure 7.

Data engineers with cloud skills are in hot demand but short supply

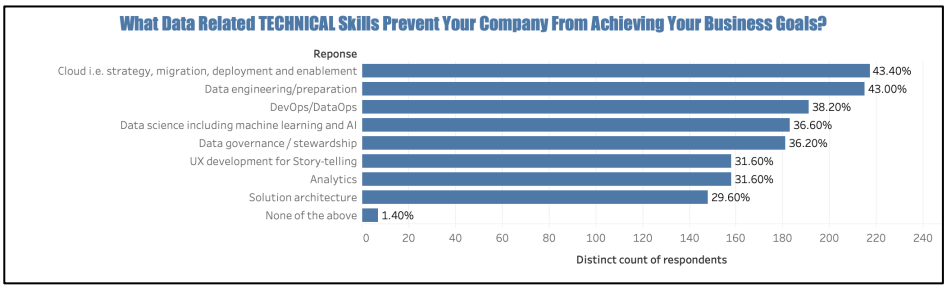

Figure 7

With respect to implementing artificial intelligence (AI), 49.2% of companies said that the biggest challenge was algorithm bias due to poor data quality (Figure 8). Also, 44.6% said that replacing current employees with AI-driven automation was a major issue.

Algorithmic bias caused by poor data quality is proving to be a major issue when implementing AI

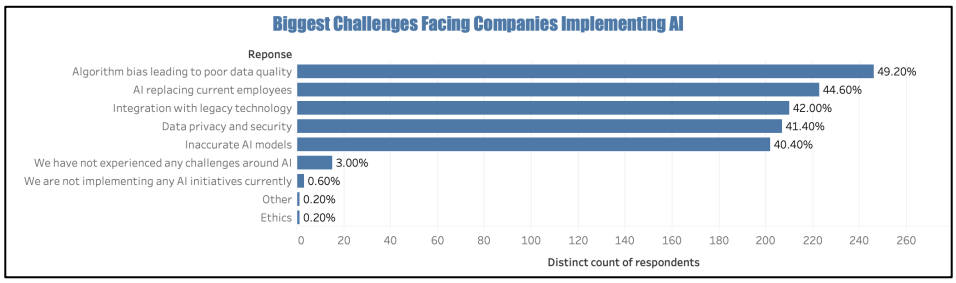

Figure 8

WHAT ARE THE MAIN BUSINESS DRIVERS?

The main business driver that stands out above everything else as to why companies want to become data-driven enterprise is the need to improve customer experience. 39.8% of companies surveyed cited this (Figure 9).

Improving customer experience is the dominant business driver for becoming a data-driven enterprise

This shows real focus from companies on keeping customers happy as digital channels such as mobile devices/apps, the company web site, and corporate social network pages start to dominate customer interaction. In these channels, customers interact with applications and not with employees. Therefore, *it is the application* that must become customer intelligent. The concern in this digital economy is that loyalty is cheap and therefore if improving and personalizing the customer experience doesn't happen customers can easily churn with the touch of a mobile phone screen or the click of a mouse.

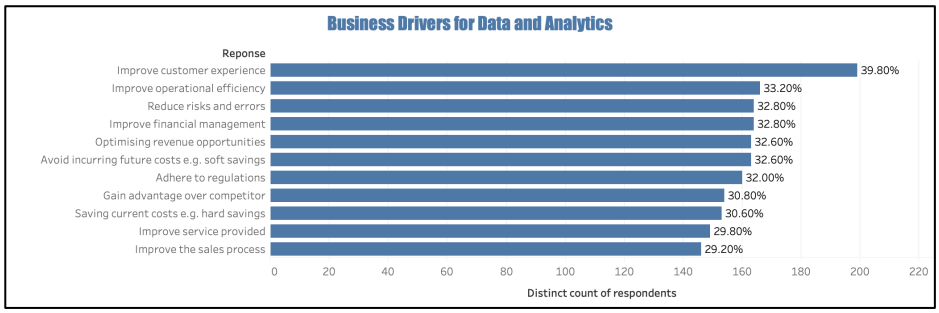

Figure 9

MAIN PRIORITIES WHEN CREATING A DATA DRIVEN ENTERPRISE

Given the challenges and customer oriented key business drivers, the main priorities companies have when becoming data-driven are very much in line with the critical success factors discussed earlier in this paper. 41.2% of companies surveyed said that data quality was their number one priority followed by the need to establish data governance (37.8%) and data tools (35%) – see Figure 10. Data skills and productivity were also deemed important.

Data quality, data governance, and data tools are higher priority than analytics

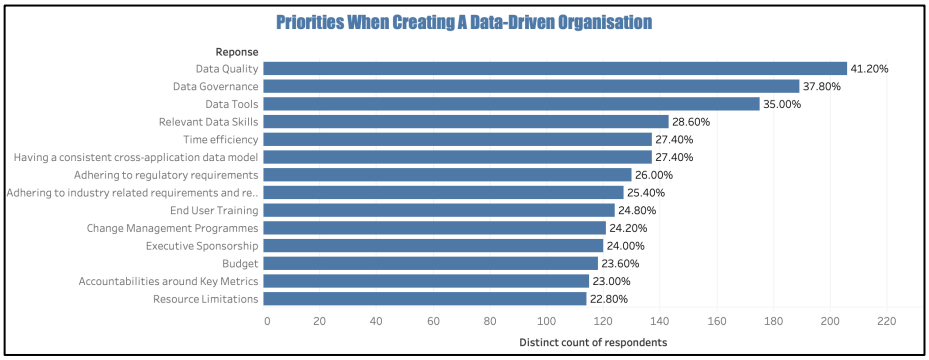

Figure 10

WHAT PROGRESS ARE THEY MAKING?

In terms of progress, most organizations are establishing a data strategy that is aligned with business strategy in order to produce the data assets needed to achieve highest priority desired outcomes.

MODERN DATA ARCHITECTURE

A key part of that data strategy is to create a modern data architecture that shows how data flows to produce high value data sets. An example of such an architecture is shown in Figure 11.

A modern data architecture in a data-driven enterprise includes managed data ingestion, a data lake, a common data curation process, a data catalog, and data virtualization software

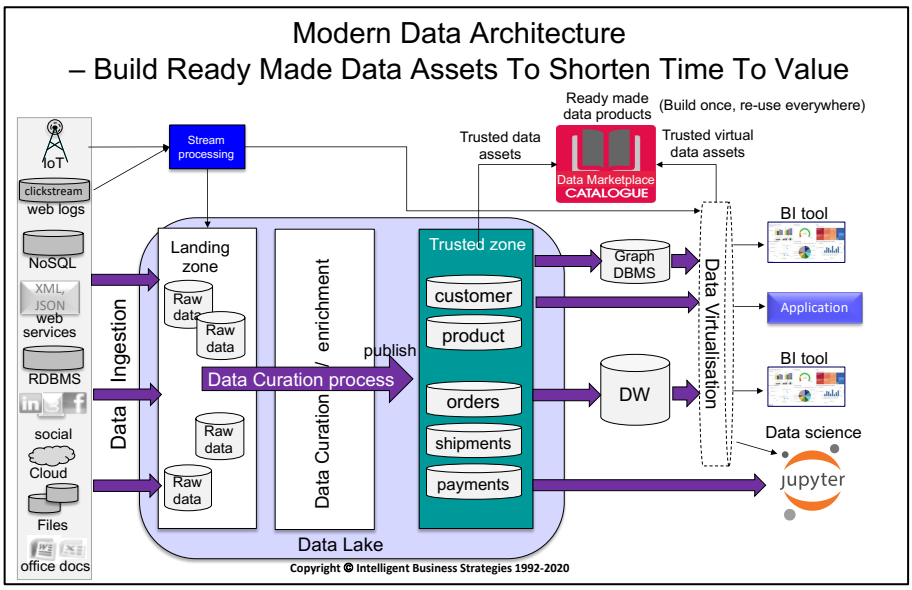

Figure 11

A data lake can be a central data store or an organized group of data stores

This shows several key components. First, the data lake is a prominent piece of the architecture with any type of data being ingested into the data lake. Note that the data could be anywhere in a distributed data landscape, from data center all the way to the edge or even beyond that for external data. The data lake can be centralized (e.g. cloud storage, or a Hadoop system) or it can be a logical data lake. The former is a centralized data store while the latter is a group of data stores organized into logical zones such that each zone contains data stores dedicated to a specific function of the data lake, e.g., they only hold raw ingested data or they only hold trusted ready-made data. Either way, the data lake is organized into zones and data tools such as data fabric software and data virtualization software are used to connect, prepare, integrate, and provision trusted data assets in a data marketplace (a data catalog). The idea is "build once, reuse everywhere." Therefore, ready-made trusted data assets can be easily found, provisioned, consumed, and used to populate a master data management system, a data warehouse, a graph database (for graph analytics), or to provide trusted data to a data scientist to train a predictive model.

Build data assets once and reuse everywhere

Publish data assets in a data marketplace catalog for others to find and use to drive vallue

Provision trusted data and simplify access using data virtualization

Data virtualization plays a key role in this architecture in terms of provisioning trusted data and also to simplify access to integrated insights across multiple analytical data stores in a logical data warehouse.

CLOUD

There is a major push to migrate analytical workloads to the cloud

In addition, many companies are building this on the cloud. In fact, 51.8% of enterprises surveyed said they would be migrating analytical workloads to the cloud including both machine learning and existing data warehouses. Only financial services and insurance are putting equal priority into migrating both operational and analytical workloads. All other industries are focused more on analytical workloads, with 35.8% of all companies surveyed saying that 41%-60% of their analytical workloads will be running on the cloud by the end of 2022. Modern data architecture component technologies like data virtualization can significantly ease analytical workload migration to the cloud. An example of this as shown in Figure 12, where data virtualization can be used to shield users from needing to know about a data warehouse migration while also enabling the data warehouse to be modernized by replacing physical data marts with virtual ones.

Data virtualization can shield business users from needing to know about data warehouse migration to the cloud

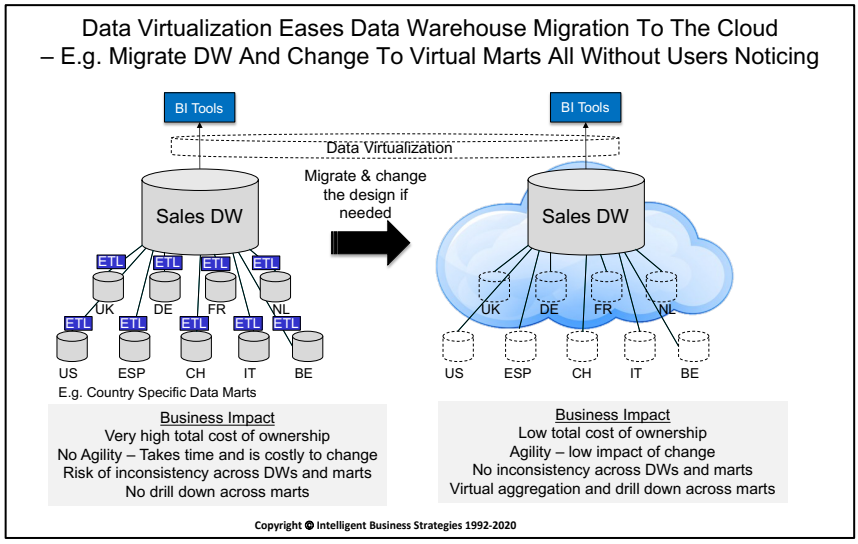

Figure 12

Data virtualization will be a key data fabric software in companies that need to operate in a hybrid cloud coumputing environment

Furthermore, in many organizations, data virtualization will be key technology in a hybrid cloud architecture, as not all workloads will be moved to the cloud.

ANALYTICS

Many companies have implemented descriptive analytics, and over 30% have implemented diagnostic analytics

In terms of analytics, there are four broad categories of analytics that represent levels of analytical progress to deliver increasing levels of business value. The lowest level is descriptive analytics with prescriptive analytics being the highest.

- Descriptive analytics - What happened?
- Diagnostic analytics - Why did it happen?
- Predictive analytics - What will happen?
- Prescriptive analytics - What should we do?

Over a quarter of companies have now implemented predictive analytics

Insurance leads the way in implementing prescriptive analytics

30.4% of companies across all industries surveyed are or have implemented diagnostic analytics while 28.4% have implemented predictive analytics. Insurance (27.78%) is the leading industry on prescriptive analytics to automate pricing in quote management, applications for insurance and claims processing whereas retail and distribution (11.11%) is still in the early stages. Utilities (39.44%) and retail (37.5%) have made significant effort in implementing predictive analytics. Also, utilities and insurance are putting more into predictive and prescriptive analytics than they are into traditional BI, which is represented more by descriptive and diagnostic analytics.

What Are the Plans Regarding People, Process, and Technologies?

In terms of planning, initiatives are underway with respect to people, process, and technology.

People

Companies are training people on data engineering and data visualization

The key trend with people is upskilling, with the most important skills needed being in data engineering and data visualization. In the interim, contract staff are being used to fill the gap.

Process

Common processes are planned for data curation and data governance

In terms of process, the focus is on establishing a common DataOps curation process utilizing component-based development. Also, data governance processes (including data quality) with automation to improve productivity (27.4% of organizations want to improve time efficiency) and establishing a consistent cross-system data model (common vocabulary) to easily share data.

Technology

The top technology that companies said they planned to use was data lakes (42.5%) followed by predictive analytics and in-memory databases (Figure 13). This clearly shows that data lakes are considered a critical part of a modern data architecture shown earlier in Figure 11.

The top technology that companies are planning to implement is a data lake

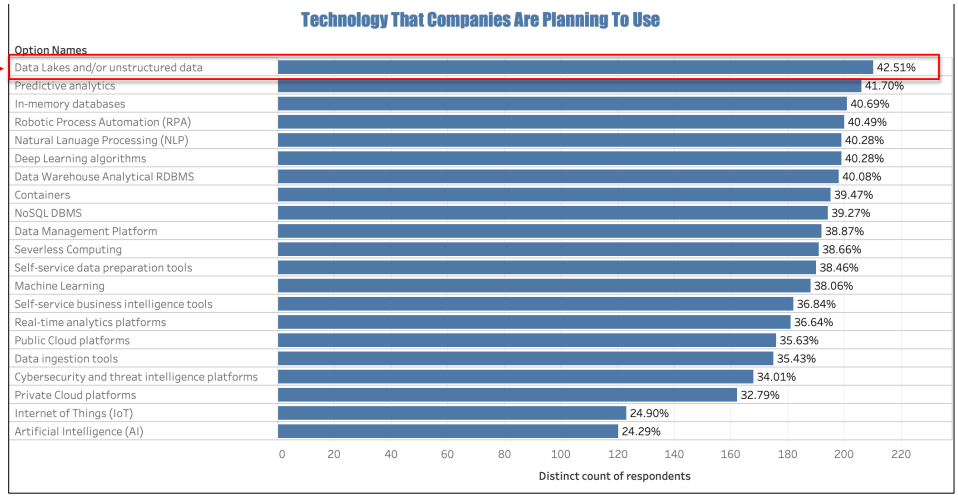

Figure 13

Data lakes are planned because people want to manage and organize data to engineer it for mulltiple use cases

It also shows a need to manage and organize data in order to ready it for use in data science, data warehousing, master data management, and other analytical workloads. Key technologies in a data lake include a data catalog, managed data ingestion, data fabric software, advanced analytics, data virtualization, and information protection technologies for security, privacy, and audit.

Many companies want to use the cloud for analytic workloads

In addition to data lakes, many organizations are planning cloud adoption for analytical workloads. This applies to all vertical industries with pharmaceutical (40.84%) and utilities (39.44%) companies pursuing this most aggressively.

WHY IS DATA VIRTUALIZATION CENTRAL TO PROVISIONING AND ACCESSING DATA IN A DATA DRIVEN ENTERPRISE?

Data-driven companies need to simplify access to data across a distributed data landscape and easily provision ready-made trusted data

Data virtualization is able to support both requirements

A key part of becoming data-driven is to establish a modern data architecture that enables organizations to produce trusted, re-usable datasets. An example of this is shown in Figure 11. This architecture should address the challenges discussed earlier, including the ability to:

- Connect to and simplify access to a wide range of data sources regardless of where data is located in a distributed data landscape
- Easily provision trusted data to whoever or whatever needs it as and when required to maximize reuse

LOGICAL DATA LAKE

Data virtualization also enables access to data even if you can't move it for legal reasons

Both of these requirements can be addressed using data virtualization, as shown in Figure 14. It can simplify access to data in multiple underlying data sources and can facilitate access to source data even though the data may be too big to move or can't be moved for legal reasons. It can also simplify access to data in multiple data stores in a data lake zone, e.g., the logical ingestion zone. In that sense, data virtualization is an essential component of a logical data lake and data fabric.

Data virtualization can provision multiple views of trusted data assets virtually without the need to copy it

It can also combine multiple data assets in a single view

The second use of data virtualization in the context of a data lake is to enable trusted physical data assets to be provisioned virtually. This avoids the need to distribute multiple physical copies of the same data to support multiple different use cases, improving consistency everywhere. Furthermore, trusted, ready-made physical datasets can be integrated at run-time to create virtual data assets that themselves can be published in a data catalog containing ready-made data assets. The result is that higher-value *virtual* data assets can be quickly built on top of trusted physical data assets to simplify access and further shorten time-to-value.

Newly created virtual views of trusted data assets data assets can themselves be published in a data catalog to create reusable virtual data assets

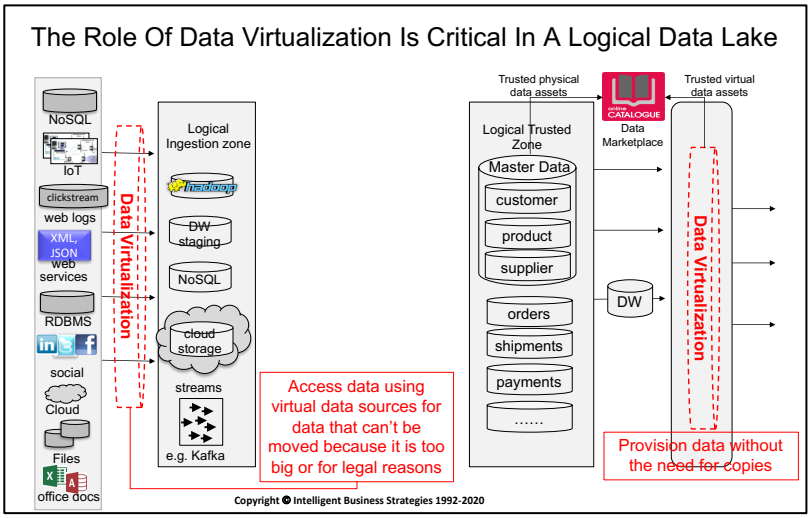

Figure 14

VIRTUAL DATA MARTS IN A MODERN DATA WAREHOUSE

Data virtualization can also be used to modernize a data warehouse by replacing physical data marts with virtual data marts

Looking back at Figure 11, data virtualization can also be used to modernize a data warehouse. One way in which this can be achieved is by replacing physical data marts with virtual data marts. This helps to significantly reduce the cost of ownership of a data warehouse by eliminating the need for physical data mart databases and data warehouse to data mart ETL processing. All of this is replaced by virtual star schema data marts that can cache frequently accessed data for better multi-dimensional query performance.

LOGICAL DATA WAREHOUSE

Data virtualization can combine data and insights from multiple analytical data stores into single a virtual view to simplify access and shorten time-to-value

In addition, data virtualization enables new insights created in different underlying analytical data stores to be presented as if it was all in one database using virtual views. This idea is known as a logical data warehouse, whereby data in a traditional data warehouse is combined with insights from other data stores including a Hadoop system, cloud storage, a graph database, or a master data management system. Even new insights produced from streaming data analytics can be included. Therefore, a complete view of a customer together with all of that customers' interactions, their relationships, and opinions could all be brought together in a logical data warehouse for use in all front-office applications (see Figure 15).

Data virtualization can be used to create a logical data warehouse that combines traditional historical data with big data

Data virtualization can be used to create a 360-degree view of a customer for use by all front-office channels

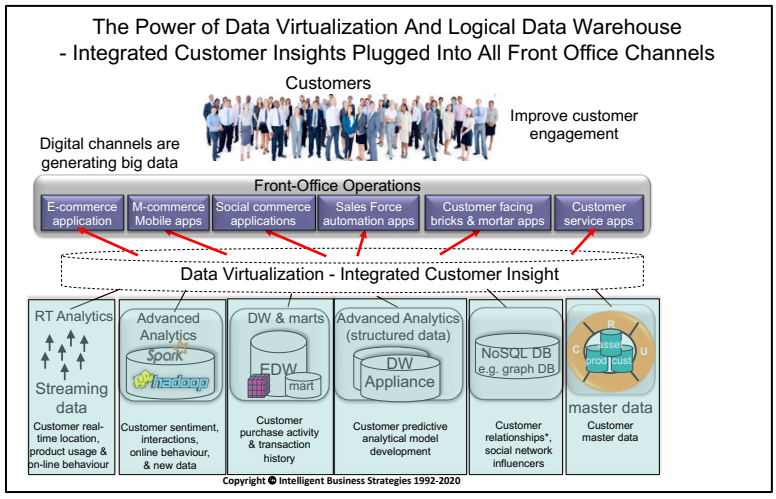

Figure 15

INTEGRATE A MODERN DATA WAREHOUSE WITH A DATA LAKE

Data virtualization can be used to integrate a data warehouse with data in an underlying data lake

It is also possible to use data virtualization within a data warehouse to enable access to data in an underlying data lake. For example, virtual views can be created to join historical data in a data warehouse to data in an underlying data lake. This means that as soon as data lands in the data lake it could be accessed and joined to historical data in a warehouse, enabling capabilities like a real-time data warehouse. Furthermore, popular virtual views can then be persisted in the data warehouse if needs be, to enrich the data there.

Denodo is a leading provider of data virtualization platform technology, which is key data fabric software needed when building a data-driven enterprise

Given this kind of flexibility, data virtualization software such as the Denodo Platform is a key technology in a modern architecture. It can handle the complexity of a distributed data landscape and can also be deployed in the data center and the cloud. In addition, the Denodo Platform also has a data catalog that provides metadata lineage of virtual views and saved queries back to the underlying data stores in the distributed data landscape.

CONCLUSIONS

Ungoverned self-service data preparation on data across a distributed data landscape will result in inconsistent data, mass re-invention and chaos in the enterprise

There is no question that the days of every department buying their own self-service data preparation tools and giving them to every business user to clean and integrate their own data are over. The trends indicate that companies do not want a "wild west" environment as described in Figure 2. What they want is to upskill their employees to improve data skills and invest in data quality, data governance, and common integrated data tools, and to organize to become data-driven.

Organising to become data-driven is the most difficult part but is critical to success

The hardest part is organizing; this requires a program office to be created, with a control board that insists on projects across the business being associated with achieving a strategic business objective before they get funded. The key point here is to start with business value and not with data. Once the value is identified, then project approval takes place such that multiple project teams will begin to create and build up trusted, reusable data and analytical assets that enable the business to achieve its goals. At present, the main business driver is to improve customer experience.

A data-driven organization needs a modern data architecture that can work across data center, multiple clouds and edge

This is happening even though different project teams may report into different managers across different departments. In other words, it is a federated organization backed up by a center of excellence to support these teams, whose job it is to co-ordinate projects and build a continuous information supply chain that is tasked with driving business value.

Data virtualization software is a key technology component in a modern data architecture to quickly provision trusted data and simplify access to insights

In addition, companies want a modern data architecture to work across the data center, multiple cloud computing environments, and the edge, and such an architecture should include a data lake, a data catalog, managed data ingestion, common data fabric software, governed data curation, published trusted data assets, and data virtualization, to quickly provision that trusted data and bring together data and insights from multiple analytical data stores to make it easy for information consumers to access, consume, use it, and act on it to drive value.

Companies who store data in a data center, one or more clouds and collect it from edge devices need data virtualization software like the Denodo Platform if they want to become data-driven

Given the trend towards an ever increasing distributed data landscape and the need for multiple analytical systems, a data virtualization platform from a vendor like Denodo is now critical to bringing a modern data architecture to life and to overcome the challenges in a data-driven enterprise. This is because data virtualization simplifies access to data across the data center, multiple cloud computing environments, and edge devices. It also enables logical data lakes to operate, even when data can't be moved for legal reasons. In addition, it can provision trusted data, virtually preventing the need to copy it, and can help modernize data warehouses by integrating them with data lakes and other analytical data stores via a logical data warehouse to drive business value based on such gains as a complete view of the customer. In fact, it is difficult to see how an enterprise can become data-driven without it.

About Intelligent Business Strategies

Intelligent Business Strategies is an independent research, education, and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, machine learning, advanced analytics, data management, big data, and enterprise business integration. Together, these technologies help an organization become an *intelligent business*.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specializes in business intelligence and enterprise business integration. With over 38 years of IT experience, Mike has consulted for dozens of companies on data strategy, data architecture, big data, machine learning, advanced analytics, data governance, master data management, and enterprise architecture. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates, an independent analyst organization. He teaches popular master classes in Data Warehouse Modernization, Designing, Managing and Operating a Multi-Purpose Data Lake, Machine Learning and Advanced Analytics, Big Data and Analytics Fundamentals, Enterprise Data Governance and Master Data Management, Modern Data Architecture and Real-Time Analytics and Operational BI.

8 Paddock Chase, Poynton
Cheshire, SK12 1XR
England
Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-mail: info@intelligentbusiness.biz