May/2021

Creating a Successful Data Fabric for Your Enterprise

Understanding the Data Fabric Processes and Technologies

Claudia Imhoff, Ph.D.

Sponsored By:

Table of Contents

Introduction

If 2020 – the year of the pandemic – taught us anything, it was that change happens and it happens suddenly, unexpectedly, and with significant impact to every aspect of our world. 2020 also taught us that, to counter these seismic changes, our enterprises must be agile. They must be able to change their entire set of objectives and goals, along with supporting operating processes and decision-making capabilities almost overnight.

For decision-making specifically, agility comes from the ability for business users to easily find and use analytical data and analytical assets for the many decisions – strategic and tactical – that must be made every day. Unfettered access to analytical data comes in the shape of a data fabric. Data fabric is a new term for an old idea – an all-encompassing analytics architecture that is easily understood and accessed, easily implemented and maintained, and ensures that all analytic data can flow easily throughout the entire enterprise.

A tall order but, thankfully, with the advances in data management, storage, analysis, and access technologies, the creation of a data fabric is not only possible, but it has become mandatory for enterprises in today's post-pandemic marketplaces.

This document begins by defining what a data fabric is, along with its benefits. The Extended Analytics Architecture (XAA) is used to illustrate the analytical components within the fabric, followed by a diagram of how data must flow (either physically or virtually) from one component to another seamlessly, thus supporting access to and the creation of the ultimate analytical assets.

The next section walks the reader through the set of processes or capabilities needed to give the user a sane and rational way to access the environment, understand and create needed analytical assets, and the ability to quickly make decisions with confidence. Within each process, the technological functionality needed to create and maintain a fully functioning data fabric environment is discussed.

The last section of the paper discusses things to think about when implementing a data fabric for all analytical needs. Using the proper set of technology can mitigate, if not eliminate, many of these challenges.

What is a Data Fabric?

Over the years, there have been multiple architectures developed in support of data analysis. The most popular ones were the Corporate Information Factory[1] and the Data Warehouse Bus Architecture[2], both created and used in the 1990's. For years, these satisfied the needs of the enterprise for analytical data when constructing an enterprise data warehouse (EDW).

But technology and time marched forward and these architectures no longer satisfied *ALL* the data requirements for analytical capabilities. Two other forms of analyses came into being – those needed for the data science community (a more fluid repository called the data lake) and real-time streaming analytics on real-time data (operational analytics). These have unique analytical characteristics and uses that simply cannot be supported within the EDW environment alone.

Enter the Data Fabric concept. The idea of a "Data Fabric" started in the early 2010's. Forrester first used the term in their published research in 2013. Since then, many papers, vendors, and analyst firms have adopted the term. The goal was to create an architecture that encompassed all forms of analytical data for any type of analysis with seamless accessibility and shareability by all those with a need for it.

Benefits of the Data Fabric

There are many benefits to a fully consolidated analytics architecture. Here are just a few:

- Easier data management, security, reliability and consistency. Well documented metadata makes the overall environment simpler to maintain.
- Democratization of data and analytical assets. This consists of easy discovery and navigation of all the data assets by all users from a centralized data access mechanism.
- Reduction of complexity. This is accomplished by promoting automation and streamlining the processes that create and maintain the environment.

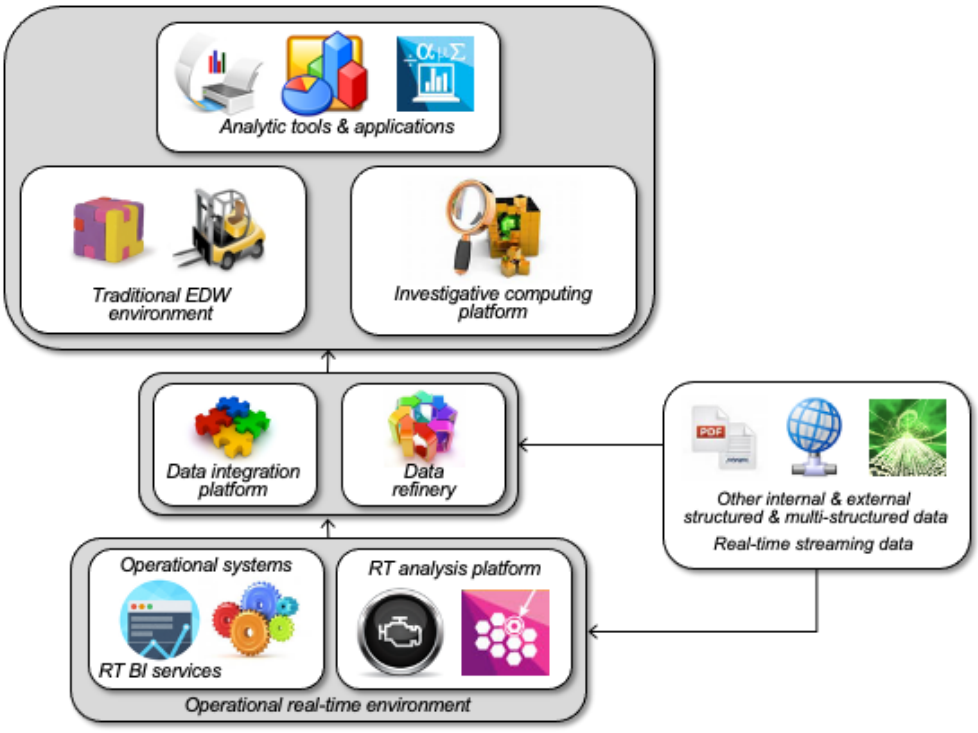
[1] "The Corporate Information Factory, 2nd Edition", by William H. Inmon, Claudia Imhoff, and Ryan Sousa, ISBN:978-0-471-39961-2, Publisher John Wiley & Sons, Inc.
[2] "The Data Warehouse Tool Kit, 3rd Edition", by Ralph Kimball and Margie Ross, ISBN: 978-1118530801, Publisher John Wiley & Sons, Inc.

- A coordinated, documented process of data lineage and usage. The more information about where data came from, what happened to it on its journey, the ultimate analytical assets that use it, and who uses that data and assets means the better managed the environment becomes. Data redundancy, inaccuracy, senescence, and potential security/privacy breaches can be better controlled here than in a chaotic, undocumented situation.

The Extended Analytics Architecture (XAA)

This author's take on the Data Fabric idea was published in 2016 as the Extended Analytics Architecture[3] (XAA – see Figure 1). Like previous architectures, this version of the Data Fabric is also a logical architecture. How an enterprise physically implements the XAA is entirely up to the enterprise, its technological resources, its personnel, and its analytical needs. The goal continues to be to create an integrated analytics environment that eliminates silos of data and analytics, supports easy access and sharing of data and analyses, and streamlines the creation and maintenance of a complex environment.

Figure 1: The Extended Analytics Architecture



[3] "Extending the Traditional Data Warehouse", by Claudia Imhoff, published in TDWI's Upside Magazine, https://tdwi.org/articles/2016/03/15/extending-traditional-data-warehouse.aspx

To understand the XAA and the Data Fabric in general, here are brief descriptions for each component, starting from the bottom:

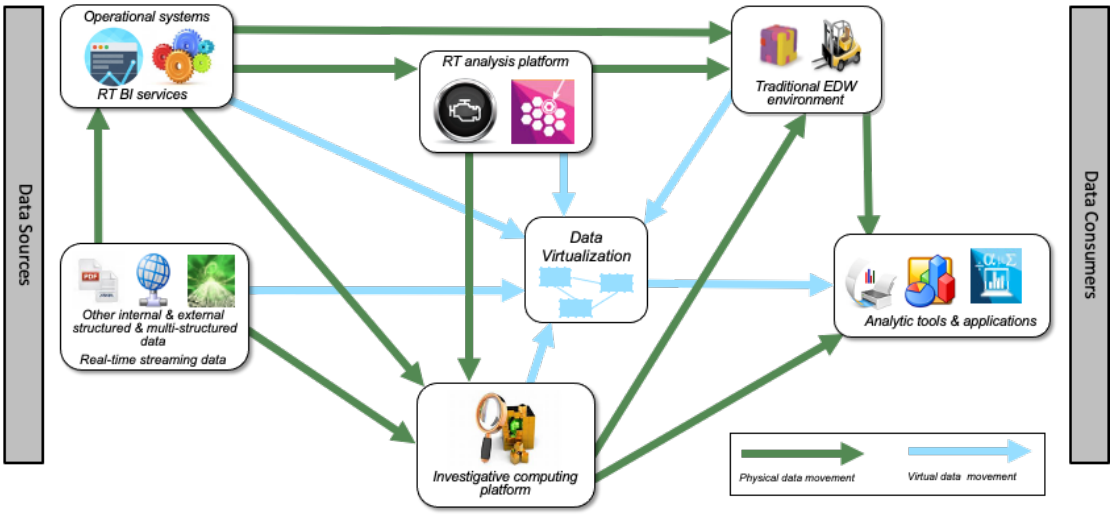
1. Operational systems – the internal applications that run the day-to-day operations of the enterprise. Within these systems are callable BI Services (Embedded BI) generally called from the EDW. Examples include fraud detection, location-based offers, and contact center optimizations.
2. Real-time analysis platform – the first analytical component that analyzes streams of data (transactions, IoT streams, etc.) coming into the enterprise in real time. Examples include traffic flow optimizations, web event analyses, other correlations of unrelated data streams (e.g., weather effects on campaigns).
3. Other internal and external structured and multi-structured data – sources of data that are not in the normal streams for this architecture and include IoT data, social media data, and purchased data.
4. Data integration platform – the process of extracting, transforming to a standard format, and loading structured data (ETL or ELT) into the EDW. The process also invokes data quality processing where needed and is considered the "trusted" data stored in the EDW.
5. Data refinery – the process of ingesting raw structured and multi-structured data, distilling it into useful formats in the Investigative computing platform for advanced analyses by data scientists. It is also called data prep or "data wrangling".
6. Traditional EDW – the second analytical component where routine, analyses, reports, KPIs, customer analyses, etc., are produced on a regular basis. The EDW is considered the production analytics environment using trusted, reliable data.
7. Investigative computing platform (ICP) – the third analytical component used for data exploration, data mining, modeling, cause and effect analyses, and general, unplanned investigations of data. It is also known as the data lake and is the "playground" of data scientists and others having unknown or unexpected queries.
8. Analytic tools and applications – the variety of technologies that create the reports, perform the analyses, display the results, and increase the productivity of analysts and data scientists.

Data Flows in the Data Fabric

For a Data Fabric to reach its full potential, the data flows within the architecture and between the various components must be thoroughly

understood. Figure 2 depicts the two forms of data flows: physical flows and virtual ones. The physical flows come from the data integration platform and the data refinery and deposit their data into either the EDW or the ICP. The physically separated data and analytical assets can be brought together to appear as if they were physically together for analysis or sharing by data virtualization. In this case though, no data is actually moved.

Figure 2: Data Flows in the Data Fabric



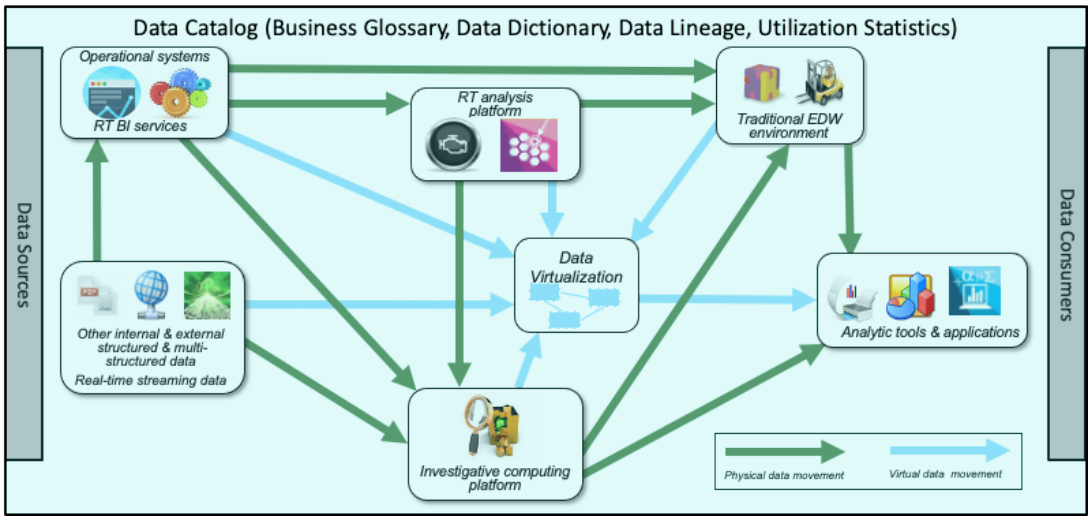
Data Catalog – Accessing and Monitoring the Data Fabric

The final technological component in a Data Fabric is an overarching repository (See Figure 3) that captures all the metadata (technical, business, and administrative) and utilization statistics within the Data Fabric. This component enables discovery: the ability to find, understand and access the contents of the Data Fabric environment. With this discovery component comes the ability to monitor the location and usage metrics of the entire environment's contents as well as enforce all data governance policies for security and privacy. This component is supported by the data catalog functionality.

A data catalog collects information about the entire Data Fabric, and acts as the gateway or entry point for the entire business community to find new data contents and existing analytical assets, to receive suggestions to help in their discovery process, and to access and analyze the data using their analytical technology. It is also where technical implementers go to understand the lineage of data (from sources to ultimate analytical assets), to receive the statistics of usage, and to begin planning any new additions or changes to the overall environment using precise impact analysis. Finally, it is where they can discover

redundancies in data and analytical assets, inefficiencies in data integration programs, and other data management deficiencies.

Figure 3: Data Catalog Component of the Data Fabric

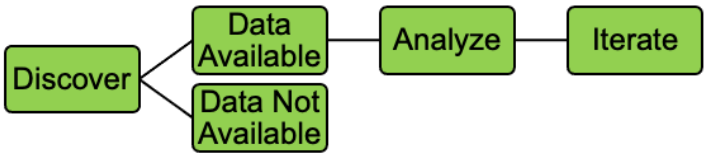


Data Fabric Processes

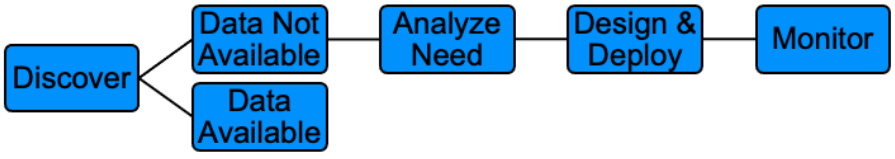
With a solid foundation in terms of the architectural components of the Data Fabric, let's dive into the processes that bring it to life. These processes will be divided into two sets – one based on the business community and the other on the technical implementation and maintenance personnel. See Figure 4 for these two sets of processes.

Figure 4: Data Fabric Processes by Role

- Business Processes



- Technical Processes



Business Processes in the Data Fabric

Most business community members are not particularly interested in the technical underpinnings of their analytical environments. Their ideal

situation is one in which they can quickly and easily access the data they need and then analyze it. How it is integrated and where it is stored may not be of great interest. Their usage of the environment should be made as easy and uncomplicated as possible.

Their processes for using the environment consist of the following:

1. Discover – The business community needs access to a data catalog function that is an easy-to-use entry point where they can quickly ascertain what data is available, in unambiguous, non-technical language that is relevant to the business, and what analytical assets exist (reports, visualizations, dashboard widgets, advanced predictive and other models, etc.). The catalog should supply browsable ontologies that give this community the context of all data and assets. It should then show the assets that contain the data with useful information such as who created the data or asset, how current it is, what its quality measurements are, what source supplied it, what technology created the asset, and so on. The data catalog should include the meaning of business rules and explanations of security and privacy policies. And finally, the catalog should be able to recommend other data or assets similar to the requested one(s) much like a cohort recommendation would ("people like you also liked these items…").

2. Data Availability – Again from the catalog, a business user can quickly determine if the data is available for their decision-making or not. If not available, then they must submit a request to the technical personnel to bring that data into the environment. If the need is critical, the technical personnel may create a *temporary workaround* (e.g., a documented one-time extract or virtualized access to data) to satisfy their need until the data is made available through the proper channels (the technical process). Once the data is made available, the temporary workaround is removed.

   If the data or asset is available, then the business users must determine if they already have permission to access it or who they should contact for permission to access the needed information. If a request is made, a governance process must be in place to ensure that the request is valid, it is appropriate for the requester, and all security and privacy policies are enforced.

3. Analyze – Once the business user has permission to access the information, the next step is to use the information in a decision-making capacity. The business user may create their own analytical asset using the data or access an existing asset, perhaps tweaking it (e.g., changing certain parameters) to fit their specific needs.

Again, the business community should ensure that any new analytical asset is noted in the data catalog feature for others to discover.

4. Iterate – After completing their analysis, the business user may continue to examine data and assets in their area or find other information by returning to the catalog. If the catalog supports the invoking of the analytical technology from within the catalog itself, then the business user does not need to have the extra step of returning to the catalog – they are already there.

Technical Processes in the Data Fabric

The people responsible for building and maintaining a data fabric like the Data Fabric have a big job. The simpler they make the business community's access and utilization of the analytics environment, the more complex the infrastructure becomes. Simplicity is not always easily attained! These technical resources will need technologies that work together seamlessly to support the following processes.

1. Discover – The technical personnel start with the same process as the business community does. However, they use Discover for very different purposes. Their goal is to discover what data and analytical assets already exist in the environment to mitigate or eliminate the duplication of either resource. They also get information from the processes that generate the technical integration, data prep, and data quality metadata of the existing data. In addition, they get the full metadata on data lineage (e.g., sources, integration techniques and processes, quality metrics, etc.), usage statistics (e.g., who is using the data and how often), and impact analysis (e.g., if an integration program changes, what downstream processes, data and analytical assets are impacted.).

2. Data Availability – From the data catalog information, the technical personnel can quickly determine if the requested data exists or not. If it exists and is readily available, the technical personnel need only update the catalog (if necessary) and inform the requester of its existence. The requester must them be vetted to ensure proper authorization for access to that data is in place.

   If the data is not available, these resources must begin their examination of the potential sources of the data. This research includes an assessment of the source's quality, content, and accessibility, and suitability for the requested purpose. Documenting all this information is mandatory input into the data catalog for future usage.

3. Analyze Need – The Data Fabric has three distinctly different analytical components – the EDW, ICP, and the RT analysis engine. Each has its own purpose and unique set of data and processing characteristics. Once a source of data has been identified, the next step is to determine which of the three components will be used for housing the data, the analytical processes that will be used on it, and the business community members that will access and use it.

   For example, reports, dashboard widgets, and routine analytical assets such as KPIs, profitability calculations, and other such trusted indicators are perfect examples of the type of data and analyses that fit the EDW perfectly. These are typically used by business analysts, corporate executives, managers, and even front-line personnel.

   Unplanned, unexpected queries like "did this ever happen?", "what-if" analyses, model development, or other advanced forms of analyses often performed by data science teams are examples of the types of data and analytics developed in the ICP or data lake.

   Finally, real-time analysis on real-time data belongs solely in the streaming analytics component in the DATA FABRIC. These include risk analyses, fraud detection, traffic analyses, IoT stream analyses, etc. These are often automated processes that perform their analyses as the data streams into the operational world.

   The technical personnel are responsible for identifying which of these analytical components should be the target environment for the requested data and analysis.

4. Design & Deploy – Based upon the analysis from the previous process, the technical personnel can begin the process of populating the right component with the right data and technologies from the appropriate source of data. For the EDW, this means performing the data integration process(es), either ETL or ELT and appropriate data quality processes to ensure the data can be trusted.

   For the ICP, the data prep is much less stringent and is used to format the incoming data so that it is easily used by data scientists and their analytical technology of choice. Data quality processing may not be as stringent or even needed at all here.

   In both the EDW and the ICP though, data security and privacy are still required. In addition, sensitive data must be identified and protected by encryption or some other masking mechanisms. The design of the data management programs for these two components can become quite complex and should be automated as much as possible to ensure rapid development, proper

documentation, improved consistency and accuracy, and ease of maintenance.

Streaming analytics is a completely different environment in which the data is first analyzed and then stored. The technology for these activities is quite different from that used in the store-then-analyze components. Furthermore, thought must be given in terms of how these analyses will be integrated into the operational workflows.

5. Monitor – In this final process, the technical personnel must ensure that the data catalog is updated with the latest additions, edits, or changes made to the DATA FABRIC, its data or its analytical assets. These personnel must also continuously monitor any changes in data lineage, usage, or impact analyses performed.

   These additions improve the Discover process for both the business community and the technical personnel in terms of future utilization, projects, or proposed changes.

Implementing the Data Fabric

A data fabric like the Extended Analytics Architecture is complicated by its very depth and breadth. It consists of multiple components, data flows and processes that must all be coordinated to achieve the goal of the architecture as stated earlier in this paper: "an architecture that encompassed all forms of analytical data for any type of analysis with seamless accessibility and shareability by all those with a need for it".

Obviously, technology plays a critically important role in achieving this goal. Every DATA FABRIC process, whether business or technical user oriented, depends on having a solid technological foundation in place. The good news is that the necessary technologies exist. The bad news is that no single company has all of these packaged as a data fabric offering.

The technological functionalities needed for a data fabric architecture are provided below. Idera has many technologies that support these. And where there is a gap, the company has developed deep partnerships (more than just a logo on a web page) with suitable vendors to fill those gaps.

1. Data catalog – A "storefront" or entry point for all users of the environment that is a repository for all technical metadata (source to target), a business glossary, data dictionary, and governance attributes. Much of this information is generated by other technologies, like data modeling, data integration or data prep tools, but it is consolidated here and presented to users via the catalog

interface. The catalog can perform certain analyses like data lineage and impact analysis. Through its AI/ML capabilities, it can determine sensitive data and present recommendations to business users. Idera's ER/Studio as well as the company's partners fulfill this role of the data catalog.

2. Data modeling – Each repository in the architecture must have its schema documented through a data modeling technology. The different levels of data models (logical and physical) are used in designing the EDW and the ICP. Data models of the operational systems are used in the data integration and prep technologies as well. Data modeling technology is also key to supplying much of the information found in the data catalog, including any changes to database design, the existence of data and its location, definitions and other glossary items. It is vital that data models are connected to the Business Glossary to ensure a well-managed data catalog. Idera's ER/Studio has been around for many years and is a proven modern modeling technology that allows data architects to connect models into a data catalog.

3. Data Integration – A key factor in the creation of the first repository of analytical data (the EDW) is data integration technology, which performs the heavy lifting of extracting the data from its sources, transforming it into the single version of the data that is then loaded into the data warehouse. This ETL or ELT process is what creates the trusted data used in many production reports and analytics. A best practice here is to automate as much of this process as you can using technologies like Idera's WhereScape, which can automate the entire process from curation to design and deployment of the EDW providing complete documentation immediately.

4. Data preparation – The creation of the second repository of analytical data, the data lake or ICP, is supported by the data prep technology that extracts raw data from sources and reformats, lightly integrates, and loads it into the repository for exploration or experimentation. This repository is meant for the unplanned, general queries performed by data scientists and the like. Idera's Qubole technology is an example of such a data prep tool that automates this process and produces the much-needed documentation for those using this repository.

5. Data virtualization – The ability to bring data together virtually rather that physically moving it all around the architecture is a great boon to the analytical personnel. Access to all data, regardless of its location, is a major step toward data democratization. The ease with which data can be virtualized is also a great productivity tool for the technical personnel. It can also be used to prototype new solutions or additions to the environment.

6. Real-time analytics engine – The last area of analysis deals with analyzing data streaming into the organization and being analyzed before it is stored. This technology must handle multiple streams of data that are analyzed in a real-time mode. It is a relatively new area of analytics but is a welcomed addition to the family of analytical components in the Data Fabric. Idera's Qubole and Idera's partners are making great strides in this exciting new analytical area.

7. Data analysis and visualization – The world of data analysis and visualization technologies is massive. The analytical capabilities of these tools range from simple reporting and dashboard creation to complex and artistic data visualizations to predictive and other advanced analyses. No single technology in this category can satisfy all business community needs. Therefore, there most likely will be multiple tools – at least one in each category, needed to support all analytical needs. Idera's Aqua Data Studio can act in many capacities in this area and is a good addition to any Data Fabric toolbox.

8. Monitoring – Many of the usage statistics found in the data catalog originate from the monitoring technologies used in the EDW and ICP. The ability to monitor what data is being used and by whom gives the technical personnel great insight into the overall performance of these analytical repositories. For example, data that is rarely used may be stored in archive media, seasonal or other time-based spikes in utilization can be planned for, and data commonly used together may be cached or virtually brought together for better performance. Without the monitoring technology, the architecture can quickly descend into chaos. Both Idera's WhereScape and Qubole have monitoring capabilities for their respective analytical components.

9. Suitable data storage – Rounding out the Data Fabric environment are the various data storage technologies – the databases themselves. In the past, we had to separate the data warehouse from the investigative area because the technologies used for both were incompatible. With the separation of data storage from compute, we are now seeing situations where the data warehouse and ICP can reside in the same storage technology.

The second major decision for many organizations is where to deploy the data – on premises or in the cloud. The decision of where data resides can be quite impactful, especially if the environment is a hybrid one – some deployed in the cloud and some on premises, or some deployed on multiple vendors' clouds. If deployed in multiple locations, bringing all the data together for analysis can be quite challenging.

Wrap-Up

The Extended Analytics Architecture is an example of a Data Fabric that encompasses all forms of data, welcomes all analysts accessing the data to create the analytical assets, and supports the technical personnel responsible for its implementation and maintenance. It is a big undertaking that requires many different technologies working together in harmony. Successful environments greatly enhance the enterprise's fact-based decision making, creates a much sharper competitive stance, and improves marketing, communications, operations, and customer relationships.

Before embarking on this endeavor, here are some things to think about:

No single technology or technology company has the complete set of capabilities listed in this paper. Some come very close to having it all, but it is important to understand where there may be gaps or holes in the offerings. If that is the case, partnerships are critical. These partnerships are far more than just a handshake; they require a deep understanding of each company's offering and must interface seamlessly between the technologies.

For any architecture to work, the enterprise must stand solidly behind the initiative creating the new environment. That means the executives may have to step in when rogue attempts are made to go around the architectural standards and components. Silos created as workarounds must be eliminated when the workaround is no longer needed.

Making the business community's interfaces simple and easily understood can make it more complicated for the technical personnel creating the infrastructure supporting these interfaces.

Much of the value of the data fabric depends on the robustness of the information gathered in the data catalog. Out-of-date, stale, or just plain inaccurate metadata leads to distrust and disuse.

Finally, simply fork-lifting legacy analytic components (like an aging data warehouse) into this fabric may be expedient but may cause problems in terms of integrating it into the whole fabric. If possible, these legacy components should be reviewed and redesigned.

Many technology companies claim they can do it all. Few can. Pick a vendor that meets the most of your technological needs and has strong partnerships with others to complete the architecture.