Intelligent Solutions, Inc.

August/2017

Data Warehouse Appliances and the New World Order of Analytics

Claudia Imhoff, Ph.D.

Sponsored By:

IBM

Table of Contents

Introduction

There can be no doubt that the architecture for analytics has evolved over its 25-30 year history. Many recent innovations have had significant impacts on this architecture since the simple concept of a single repository of data called a data warehouse. First, the data warehouse appliance (DWA), along with the advent of the NoSQL revolution, self-service analytics, and other trends, has had a dramatic impact on the traditional architecture. Second, the emergence of data science, real-time operational analytics, and self-service demands has certainly had a substantial effect on the analytical architecture.

The single data warehouse repository simply could not support any and all analytics anymore. A new architectural concept called the Extended Data Warehouse architecture (XDW) has taken the place of the single repository idea. It accommodates the new forms and volumes of data, the need for different sub-environments for varying analytical requirements, and the immensely innovative technologies available today.

In this paper, we focus on the DWA and how it has evolved over the years since its introduction. The XDW architecture is then described, in which the need to maintain the data warehouse is documented while adding new components and capabilities to extend the analytical capabilities. This section also discusses the appropriate usage of appliances within the XDW. The rest of the paper covers the benefits from implementing the DWA, the selection considerations for them and what the future holds for them.

What is a Data Warehouse Appliance?

The data warehouse appliance term became popular in the early 2000’s[1]. At that time, its definition was:

“A turnkey, fully integrated stack of CPU, memory, storage, operating system (OS), and RDBMS software that is purpose-built and optimized for data warehousing and business intelligence workloads.”[2]

Fundamentally, it is an embedded, purpose-built, advanced analytics platform that enables analytic enterprises to meet and exceed their

[1] From: https://en.wikipedia.org/wiki/Data_warehouse_appliance
[2] From: http://www.infostor.com/index/articles/display/293088/articles/infostor/top-news/introducing-data-warehouse-appliances.html

business demands. Under the covers, the DWA is an integrated software and hardware bundle from a single vendor designed specifically for data warehousing and analytics.

Data warehouse appliances started life as on-premises “black boxes” that you simply loaded with your data and to which you attached your favorite BI technology. They were specifically designed for high performance analytics, delivered in an easy-to-use, preconfigured system. Truth be told, this black box characteristic did make IT rather nervous but more on that later.

These devices have evolved into expert integrated systems with built-in analytic functions and a simplified user experience. They are easy deployments requiring no tuning and minimal maintenance. They use the latest in database capabilities such as in-memory capabilities, columnar storage, MPP architectures, data skipping, and data compression to name just a few.

Today, modern DWAs provide a containerized software environment converting DWAs from the “black boxes” of days gone by into private cloud computing platforms. Private cloud platform DWAs have these benefits:

- A flexible computing platform for data lake and other flexible self-service access solutions needed by data scientists; and
- Seamless integration with analytic applications and open source tools within the data warehouse itself

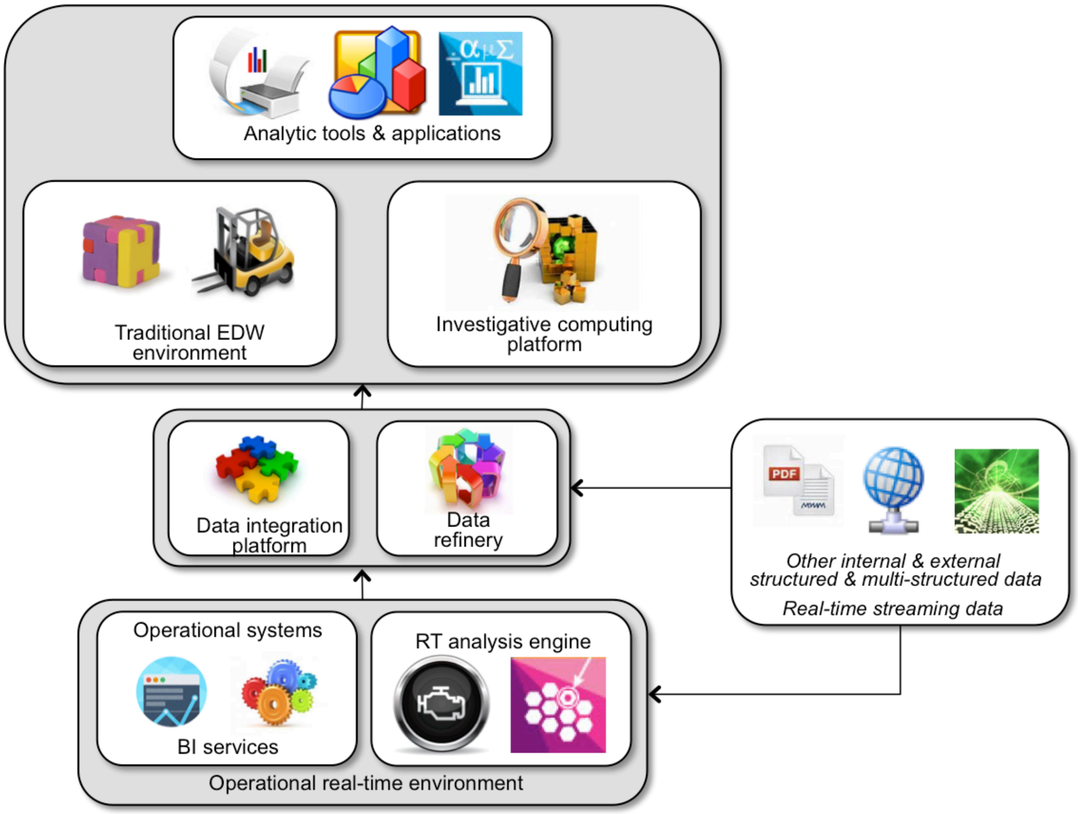
The Extended Data Warehouse: A New Architecture for Analytics

In recent years, there have been several articles, research papers, and presentations that declared: “The data warehouse is dead.” This was nothing more than marketing hyperbole, intended to get people’s attention. Unfortunately, it did, but also unfortunately, it is completely erroneous.

The data warehouse is alive and well. It’s once lofty position as the *only* source of all analytics has certainly disappeared. This leaves us with questions about what other components are needed in modern analytics architectures and what does the new architecture look like after adding in these new components.

The answer lies in the Extended Data Warehouse (XDW), a comprehensive analytical architecture that encompasses many sources of analytics, including the much-maligned data warehouse (see Figure 1).

Figure 1: The Extended Data Warehouse Architecture


Analytic tools & applications
Traditional EDW environment
Investigative computing platform
Data integration platform
Data refinery
Other internal & external structured & multi-structured data
Real-time streaming data
Operational systems
BI services
RT analysis engine
Operational real-time environment

The XDW goes by other names like logical analytical architecture, hybrid environment or modified data lake environment.

The XDW Components

Let's start with the traditional enterprise data warehouse (EDW) and answer the question; "Does the EDW still have a role in the new analytical landscape"? The answer is a resounding yes – at least for the foreseeable future of analytics. But its role has changed. It has become the source for established standard *production* reports, historical comparisons, and analytics. The data warehouse is still the best source for reliable, consistent, integrated quality data for critical or sensitive analyses, especially for financial, compliance or regulatory requirements. It is also the source for standard dashboard components (think corporate KPIs) and standard metrics like profitability measurements used by operations, marketing, sales, and other departments. Nothing beats this workhorse and its associated data integration and quality/profiling processes (Data Integration Platform) for these vital and trusted production analytical deliverables.

But it must also be acknowledged that the EDW has limitations as well – especially when dealing with unusual types and volumes of data, experimental or investigative analyses, and real-time analyses. These new requirements mandate that some analytics move outside the traditional EDW to new components in our analytics architecture.

The first new component is the Investigative Computing Platform or as others may call it, the Data Lake. This component is used for the *exploration* of big data sources and developing specialized analytics like data mining, cause and effect analyses, "what if" explorations, and pattern analysis, as well as general, unplanned investigations of this data. Some organizations may use the investigative computing platform as a simple sandbox for experimentation; others implement it as a full analytic platform or use it as an extension of the data refinery (described below). This component gives companies the ability to freely analyze and experiment with large volumes of data with phenomenal performance. Output from these activities could be used by the EDW, a real-time analysis engine in the operational environment, or within a line-of-business application (embedded BI).

The next component is a new one in the data management area – the Data Refinery. Its purpose is to ingest raw, detailed data in batch and/or in near real-time from new and unusual sources of big data (such as sensors, social media, IoT, and RFID tags). These sources are loaded it into a managed relational or non-relational data store. The data refinery – just like its counterpart, an oil refinery – *distills raw (big) data* into useful and usable information and distributes it to other components (mostly into the Investigative Computing Platform). The technologies supporting data preparation fit into this new data management area nicely.

The final extension to the data warehouse architecture supports operational intelligence through a Real-time (RT) Analysis Platform found within the operational environment. Its purpose is to develop and/or deploy real-time analytical applications or *streaming analytics* for applications like fraud detection, web event analysis, traffic flow optimization, and risk analysis. The models and rules embedded in the RT Analysis Platform are most likely developed in the EDW or investigative computing components or within the RT Analysis Platform itself, requiring tight integration and freely flowing data to and from these components.

Where Does the Data Warehouse Appliance Fit?

The XDW is a logical architecture; how it is physically implemented is up to the resources executing its construction. For example, the team may use a data warehouse appliance for the EDW, Hadoop for the investigative computing platform, and a complex event processing or event stream processing product for streaming analytics.

The good news is that today's DWA can handle more than one XDW component. A DWA, such as IBM's Integrated Analytics System and PureData for Analytics and their innovations in data storage, performance and scalability, are perfectly suited for both the EDW and the investigative computing platform. They will also work for low latency operational analytics. And with cloud implementations, the elasticity of data storage means they can certainly support both the EDW and investigative computing platform that have varying storage needs.

Streaming analytics or analytics on truly real-time data requires a different architecture from the store-then-analyze one for the EDW and investigative computing platform. Streaming real time engines analyze a stream of data first and then may store the data. Therefore, the DWA may not be suitable for these analytics – yet.

The decision to use what technology where in the XDW depends on a number of factors. Here are a few to consider:

1. Top of mind is usually the total cost of ownership. Companies with tight budgets and limited technology resources should look into DWAs. "Outsourcing" the effort of construction and maintenance of an analytics environment is very appealing.

2. Second to cost is the amount of data needed for the expected analytics. Technology selection must balance ease of data storage with maximum data scalability and performance. Again the cloud DWA offerings have a great advantage in the ability to expand or contract data storage easily and quickly depending on data and workload requirements. On-premises versions must be chosen with an eye to the future for new data requirements and so buy for those distant needs upfront.

3. The actual types of analytics must be considered. Will the environment be used for simple descriptive analytics like reports and comparison analytics? Will the environment be used for more complex and complicated analyses like predictive and prescriptive ones? Will the environment have to support production analytics as well as the unexpected and stream-of-consciousness ones from data scientists? Will there be streaming analytics as well as conventional historical ones? The analytical environment may have to support one, two or all three analytical capabilities.

4. Finally, in determining the appropriate technologies, consider the sophistication of the underlying data management functions. Modern DWAs have simplified processes for data intake. They need an in-memory optimized database or the equivalent for query performance and low maintenance (no tuning). They must have an easy administration UI and be able to handle any and all types and volumes of data. Finally, they should have built-in analytic (in-database) algorithms and models like linear regression and k-means clustering, as well as geospatial extensions.

In addition to being the EDW and investigative computing platform repositories, the DWA has other use cases that include self-service data access, dynamic workload management, ability to query archived data (especially when combined with cloud or Hadoop data stores), and data preparation (e.g., as part of a data lake component).

Benefits from Data Warehouse Appliances

We discussed a few of the benefits to a data warehouse appliance already but let’s list out a few of the more important ones here:

- An analytical architecture like the XDW is complex, having many moving parts within its domain. It can be especially confusing if implementers decide to use different vendors’ technologies or ones that are new to organization. The DWA is a substantial breakthrough in that it removes all this complexity for the purchasing entity. It reduces the many vendors involved in a traditional hardware and software combination to a single vendor, responsible for the entire stack. Bottom line: one call, one contact point.

- The preconfigured aspect of the DWA is another significant advantage. Ultimately, the idea behind a DWA is to provide a self-managing, self-tuning, plug-and-play database system that can be scaled out in a modular, cost-effective manner. There is great simplicity, no configuration, and linear scalability. Add the ability to deploy it either on-premises or in the cloud and you have a remarkably accommodating environment for all sized enterprises and analytics requirements.
- Another significant benefit for the DWA environment is the remarkable security for the data housed within it. For example, IBM's Integrated Analytics System and PureData for Analytics have automated encryption for data at rest and in transit, database activity monitoring, database access control for authorizing users and deployment hardening (behind the firewall) eliminating port scans and other network security threats.
- The on-premises DWA is a good solution in situations where analytics are quickly needed but a cloud implementation is not feasible because of regulations or privacy/security concerns. The on-premises version has all the benefits expected in an appliance – ease of installation, no tuning needed, plug-and-play functionality, scalability and performance.

There is a final benefit that is not as well known as these. It is a relatively new idea for alternative utilization of these appliances. The idea is to use it as a "containerized" environment for other, non-analytical applications. The DWA could be a suitable environment for more than analytical workloads; an organization may, in fact, load additional applications onto the DWA that are more operational in nature than analytical. For example, you could use your DWA to run your CRM system or office productivity tools in addition to your analytical requirements.

Data Warehouse Appliance Selection Considerations

Data warehouse appliances do have significant benefits but there are also some considerations as to their suitability in some organizations.

- As mentioned, some IT organizations are wary of black box technologies. The concept that you cannot tune or configure anything in the box is anathema to many DBAs and other database managers. Should something go wrong, how can they fix it? Also, they may fear that their jobs may be diminished or lost. The basic principle behind the appliance is that the vendors have deep expertise is all aspects of data warehousing. Therefore, they have configured the hardware and software to work perfectly with the data and the queries submitted to it.

- While the DWA does remove many of the technological barriers to setting up and configuring the hardware and software, it does not relieve the implementers from the often-onerous job of integrating and cleaning up the data. Sometimes you can simply load and go, especially if the analyses are approximations or experimental in nature. Other times, the data must be curated and certified as to its accuracy. The DWA is simply the data repository so ensure that the data is of the correct level of integration and accuracy for the intended users.
- Lastly, the purchaser of DWA technology should ensure that the vendor gives them the maximum flexibility when it comes to deployment options. DWA technology should support the concept of "design once, deploy anywhere". The team may start with an on-premises DWA and then decide that the cloud version would be better for the enterprise. If that is a possible future decision, the best vendor is one that uses the same technology environment regardless of where it is deployed (cloud or on-premises). This gives the purchasing enterprise easier administration (the team won't need different skills to manage different form factors) and easier workload management (the same engine makes it simple to move workloads around, with no tuning or configuration required).

The Future of Data Warehouse Appliances

Data warehouse appliances will have a role in the analytics environment for years to come. The reduction in overall TCO and power utilization combined with the simplicity and convenience of a cloud platform make for a compelling argument for its usage.

Its flexibility, scalability and security features have positioned the appliance well for supporting significant analytical components in the Extended Data Warehouse architecture. The DWA is certainly capable of supporting the traditional EDW and its production analytics as well as the investigative computing platform for experimental and exploratory big data requirements.

Because of its rapid load and go capability for massive volumes of data, its documented performance, and its ability to handle mixed workloads, it is an ideal environment to support short, simple queries to the more complex, advanced analytics. Therefore, the DWA is a suitable environment for traditional BI reporting and multi-dimensional analytics, self-service analytics for entrepreneurial users and traditional business analysts, as well as the more advanced data science professionals alike.