Ensuring a Sustainable Architecture for Data Analytics

Claudia Imhoff, Ph.D.

Table of Contents

Introduction 2

The Extended Data Warehouse Architecture 3

Integration of Three Analytic Environments 4

Building for the Future 6

Summary 8

Sponsored By IDERA Inc.

IDERA understands that IT doesn't run on the network – it runs on the data and databases that power your business. That's why we design our products with the database as the nucleus of your IT universe.

Our database lifecycle management solutions allow database and IT professionals to design, monitor and manage data systems with complete confidence, whether in the cloud or on-premises.

ER/Studio is the collaborative data modeling solution for data professionals to map and manage data and metadata for multiple platforms in a business-driven enterprise data architecture.

Whatever your need, IDERA has a solution.

Introduction

Today, you can't pick up a magazine, read a blog, or hear a webcast without the mention of Big Data. It has had a profound effect on our traditional analytical architectures, our data management functions, and of course, the analytical outputs. This paper describes an analytical architecture that expands on the existing Enterprise Data Warehouse (EDW) to include new data sources, storage mechanisms, and data handling techniques needed to support both conventional sources of data and those supplying Big Data.

The Extended Data Warehouse Architecture

The real value of a decision-making environment is not in the creation of reports or simple multi-dimensional analytics. It is in the development and use of the more sophisticated analytics like statistics, predictive algorithms and data mining using all sources of data. And the way to support these critical capabilities is by extending the traditional data warehouse environment to include data sets in more fluid, less controlled components in addition to the EDW. Everyone in the enterprise—not just highly trained data scientists or statisticians—should be able to access all architectural components to include these valuable capabilities in their everyday decision making.

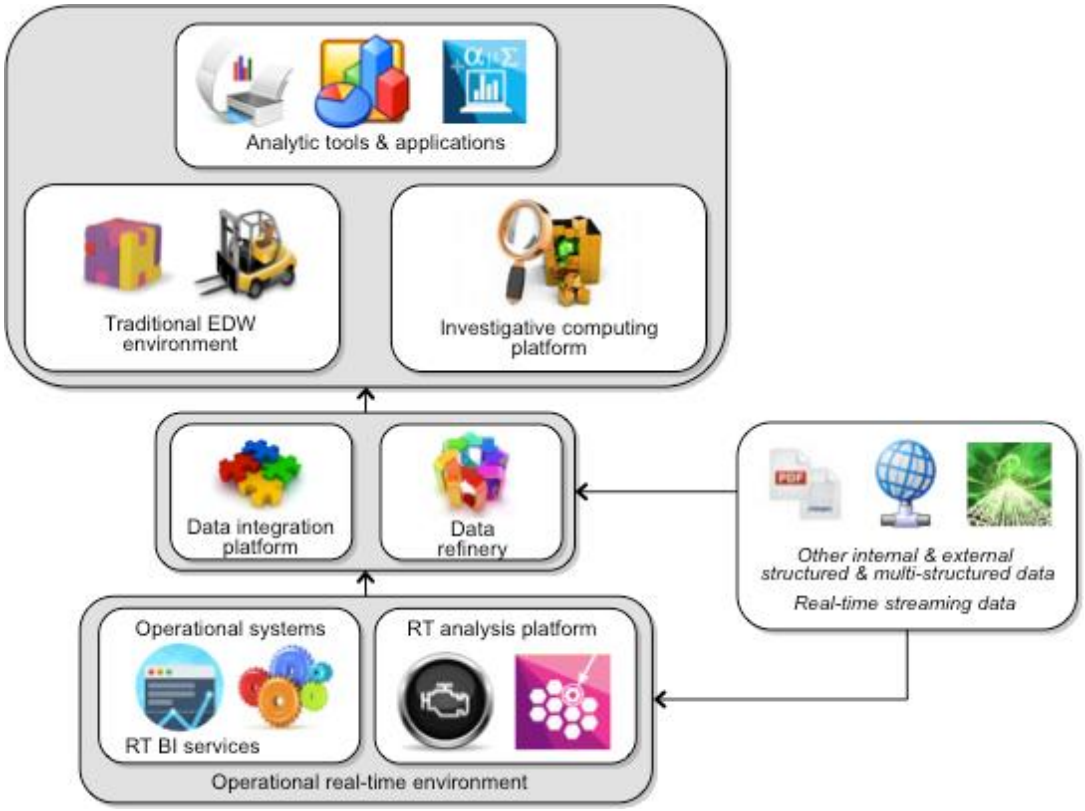
The Extended Data Warehouse Architecture or XDW (Figure 1) was created to support all forms of data, analytics and the business community creating and using these. It consists of three analytical environments that must work together to give enterprises consistent, reliable and global views of customers, products, markets, etc.

- Traditional EDW environment – we are most familiar with the components to create the EDW. They start at the bottom of the XDW with operational data, move through the data integration platform for ETL and data quality processing, and end with the loading of the integrated data into the EDW platform. This becomes the basis for *production analytics.* These standard reporting and analytics are created using the analytical tools of choice and myriad analytical applications.
- The Investigative Computing Platform – this second analytical environment is for data exploration and experimentation, mostly using non-traditional or Big Data sources. The data is brought into a data refinery (some may call it a data lake or holding pond) where it is prepared for analysis. Once prepared, it enters the Investigative

Computing Platform where unplanned, general analyses are performed using the analytical tools or applications. Note: To date, this environment is not a replacement for the EDW but rather an enhancement of the overall architecture.

- Operational intelligence – the third analytical environment performs *real-time* analytics on *real-time* data either through callable BI services (usually supplied by the EDW environment) or via a real-time (RT) analysis engine for streaming analytics. The models and rules embedded in the RT Analysis Platform are most likely developed in the EDW or investigative components or within the RT Analysis Platform itself, requiring tight integration and freely flowing data to and from these components.

Figure 1: The Extended Data Warehouse Architecture (XDW)



Integration of Three Analytic Environments

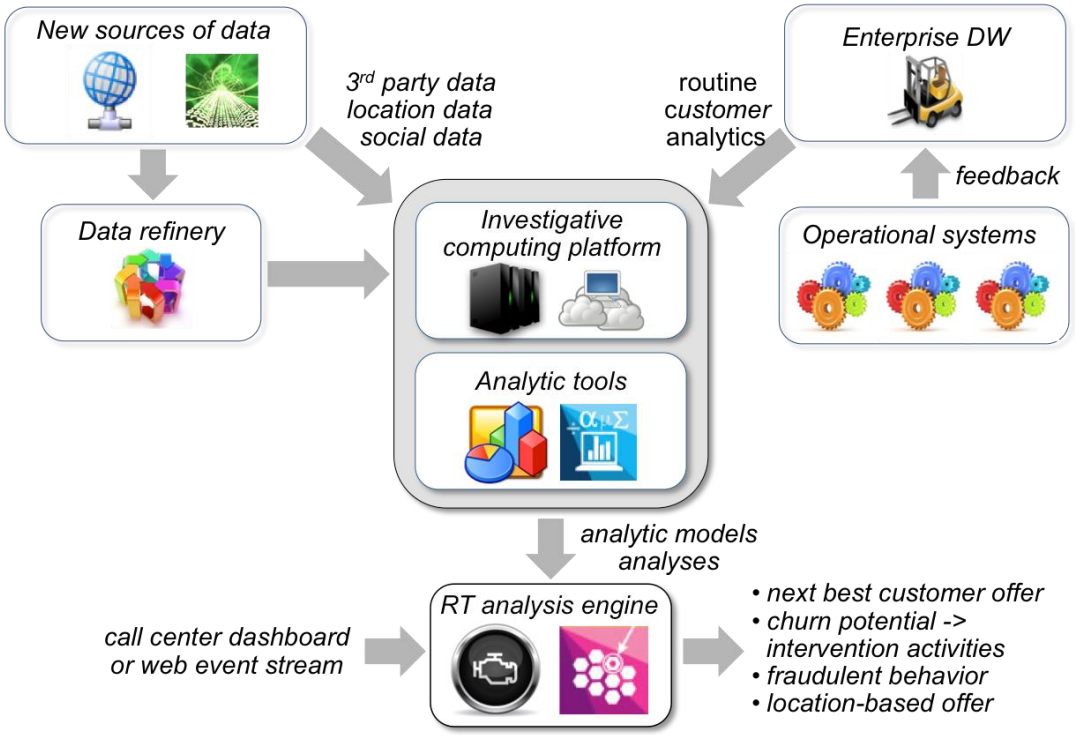
The XDW has two data integration processes in use that are often confused. The next section differentiates between a data integration platform and a data refinery:

- The Data Integration Platform performs the heavy lifting process of extracting, transforming to standard format and loading (ETL) structured data in a mostly batch fashion. This process physically consolidates data into "trusted" EDW sets for analysis and invokes data quality processing where needed. The platform employs low-cost hardware and software to enable large data volumes to be combined and stored. It also requires more formal governance policies to manage data security, privacy, quality, archiving and ultimate data destruction.
- The second data management capability is the Data Refinery. Its purpose is to ingest raw detailed data in batch and/or in real-time from new and unusual sources of big data, and load it into a managed relational or non-relational data store. The data refinery distills raw data into useful and usable information via data wrangling or "munging" (a form of "ETL-Lite") and distributes it to other components (e.g., the Investigative Computing Platform or EDW). The process also employs low-cost hardware and software to manage large amounts of detailed data cost effectively. The data refinery requires more flexible governance policies to manage data security, privacy, quality, archiving and destruction.

As mentioned earlier, enterprises are implementing these environments to give them fuller, richer analytics about their critical entities. However, the three environments cannot be implemented in isolation from each other; the analytics must ultimately be brought together physically or, more likely, virtually, and presented to the business community for decision-making. You will find that data virtualization becomes another critical data integration tool for the XDW.

As an example, Figure 2 shows the various analytics working together – in this case virtually – to present the results to a call center representative or a visitor to the enterprise's website. The EDW produces routine customer analytics like lifetime value scores, segmentation, buying behaviors, and propensity to purchase. External data flows in with more customer data like their current location and social media posts. These may flow directly into the investigative computing platform or through the refinery first before further analysis like sentiment analysis. All these analytical results are used to create the various analytical models and other advanced analyses, which are fed into the RT analysis engine, ultimately creating a next-best product offer, intervention activities based on churn likelihood, or location-based offers to customers. To bring all these analytical results together, many companies use virtualization technologies for a more flexible, mix-and-match approach.

Figure 2: All XDW components must work together to create actionable analytics.



Building for the Future

There can be no doubt that the past few years have been a period of incredible innovation in analytical technologies. Unfortunately, with significant innovation comes significant disruption. The XDW is a result of that disruption. It puts all the components into perspective – each with their appropriate roles and features but the questions still come up: "How do I get started in this new world order and how do I future-proof what I build today"? Here are some suggestions that should help ensure a successful transition.

- Create an analytical architecture. Whether you start with the XDW and modify it for your needs or create your own from scratch, you must have a blueprint or architecture to serve as a roadmap for current and future project implementations. The architecture guides everything from platform selections to performance requirements to ultimate deployment options (e.g., on-premises, cloud, self-service).
- Perform an audit of your current capabilities. No one wants to throw the baby out with the bath water, so study your current technologies and analytical capabilities. Determine the gaps from where you are now to where you want to be. Then determine what new or different

technologies will be needed to support these future requirements. Remember: the data warehouse remains a viable and valuable component of any analytical environment but it cannot support all forms of analytics.

- A data model is a must-have. The architecture may be the blueprint but the data model documents the intricate wiring (data and relationships), information (metadata) about critical components, and the maintenance of their environments (governance). These make any environment "future-proof", eliminating the need to start from scratch. Let's face it: metadata does not change as rapidly as the underlying technology does. Business glossaries are critical for sustaining the data models regardless of the technological choices made. Basically, the creation of these pieces of enterprise intellectual property means the technologies and even the processes can change and be replaced without the loss of the corporate memory.
- Ensure that all members of the business community can easily understand analytics. Data science is the darling of the analytical world today but these resources make up a very small portion of the overall business user population. Data visualization and easier user interfaces help the more technologically-naïve users access and consume analytical results for better decision-making. However, the enterprise should not get so overly enamored with data visualizations that it forgets what business problems it is trying to solve with them.
- Perform education on analytical thinking for the entire business community. Everyone in an enterprise is an analyst at certain points during their day and they must have access to analytics and make decisions using this intelligence. But do note: education is NOT the same as training on the various analytical technologies. It includes critical thinking, which is the ability to apply reasoning and logic to new and/or unfamiliar ideas, views, and situations. Basically it is educating people to ask the next logical questions in a series. For example, what happened (descriptive analytics), why did it happen (diagnostic analytics), will it happen again or continue on this path (predictive analytics), and finally, what should I do (prescriptive analytics)?
- IT must have monitoring and oversight into the analytics environment. IT cannot control the information assets used by the business, nor should it try. However, it should have the ability to scrutinize its usage for privacy and security breaches as well as for

analytic results no longer needed or used. It should have supervision over technological performance to detect bottlenecks or unusual activities as well as governance over sensitive data or analytic results. For example, it is often difficult to determine what data must be governed and what data can be ignored. If IT can monitor the usage of data, it can quickly decide that data used in regulatory or compliance reports must be governed. Data used in approximations or estimated analytics can be ignored.

Summary

Analytic environments must keep up with the technological advances and expanding business needs occurring today but that doesn't mean chaos will reign. The drivers moving enterprises to extend their data warehouse environments include:

1. The need for fast time-to-value to gain business benefits. It is impractical to use traditional data warehouse approaches for all analytical solutions. We must consider extending the environment to include real-time analysis engines, embedded BI, and investigative computing platforms.
2. The need for high performance solutions to support new analytic workloads. "One size fits all" data management is no longer viable. Enterprises must match the technologies and costs to business needs and the required analytical workloads.
3. The need to modify data modeling and integration approaches. Analytical environments need to support new data types, sources and platforms as well as new data integration approaches like data blending, data wrangling, schema-on-read, and data repositories.
4. The need to modify data governance approaches. It is no longer practical to rigidly control and govern all forms of data. Enterprises are developing different levels of governance based on security, compliance, quality and retention needs.

Thoughtful examination of current and future analytical requirements, the creation of an expanded architecture, and the development and consistent use of data models and their associated intellectual properties will significantly "future-proof" an enterprise's analytical capabilities.