published in the Proceedings of the International Conference on Computing and Data Analysis, Lakeland, FL, May 19-23, 2017

The Data Warehousing (R)Evolution: Where's it headed next?

Jeffrey Smith and Manjeet Rege
Graduate Programs in Software
University of St. Thomas
St. Paul, MN 55105, USA
{smit9169, rege}@stthomas.edu

ABSTRACT

This paper provides an overview of the history and current state of data warehousing and corporate analytics. It begins with a quick review of the history of the data warehouse and then does a deeper dive into subsets of this space including data integration, the DBMS, business intelligence & analytics, advanced analytics, and information stewardship. It finishes with a quick review of some of the leading trends in data warehousing including Big Data and the Logical Data Warehouse, Hybrid Transaction Analytical Processing and In Memory Computing.

1. INTRODUCTION

The volume of data in the modern corporate world is increasing by 35%-50% every year. On average, a typical large company today processes upwards of 60 teravbytes of data annually, which is about a 1,000 times more than a decage ago [1]. With this in mind, one can easily see the necessity of having tools and processes for managing that data. Data management really took off with the appearance of data warehousing. Some of the early contributors to the field include Bill Inmon [2] and Ralph Kimball [3]. Inmon presented the idea of the Corporate Information Factory (CIF) which later on was extended to account the stratification of data based on data currency and incorporated the critical data governance components. Kimball came up with the concept of the dimensional model, then, evolved it into conformed dimensions. His next stage brought in Master Data Management (MDM), and finally incorporated the MDM with data marts. While, there has been prior work comparing the two approaches, our focus in this work is to provide a brief background and present the current and future direction of data warehousing in general.

Back in the 1980's and early 1990's most corporations were focused on consolidating data into a warehouse for increased efficiency and data accuracy. There wasn't a tremendous amount of thought going into how it would be used, and a waterfall project management methodology was used with a timeline measured in years. Unfortunately with the large investment required in this new technology and little to show for it in the early stages, many executive management teams quickly soured on the idea and DW projects were shut down before they had delivered much of anything to the end users.

However in the mid 1990's a group of BI vendors such as Cognos, Business Objects, Hyperion, and Microstrategy entered the market with much more user friendly applications that allowed quick turnaround of report development and analytics that allowed business users to "slice and dice" data with or without a fully developed data warehouse. At the same time the DBMS vendors began adding BI type functionality to their databases with technologies like data cubes and OLAP (On Line Analytical Processing) [4]. This provided a much needed boost to the industry and reenergized spending in this area. The development life cycle was typically shortened to three to six month deliverables, which was much more palatable for executive management.

As BI and Data Warehousing finally got firmly rooted in the corporate culture, the focus split into two different directions. There still was the need to develop more efficient management of all this data and we needed to be able to handle the new technologies and types of data that continued to present themselves. In the area of Data Warehousing, new DBMS vendors such as Teradata, and more recently Netezza, began to establish themselves with different architectures that better handled analytics. Data modeling concepts such as Operational Data Stores (ODS), logical data warehouses, along with a myriad of data mart concepts found their place into the model. These concepts, coupled with advances in ETL and Middleware, allowed for much greater efficiency of data movement and data processing. The concept of metadata and Master Data Management (MDM) also started to become critical. There was a need to be able to track all the data, knowing where it was coming from, where it was going, how it was being changed, and how it was being used. An entire new discipline with tools and roles rose up around data governance just to keep track of and manage it all. Additionally, near real time analytics and dashboarding started to become much more important to the organization. No longer was it good enough to know how manufacturing output or sales were last quarter or last month. With the rise of the internet and faster pace of information we now wanted to know how we did in the last hour or even last five minutes. This became especially important as data such as manufacturing process controls and click stream data started to get incorporated into the warehouse

and we needed an efficient way to identify and visualize the items of importance. This also led to the increased importance of data mining and data analytics tools as the volumes of data grew from Gigabytes, to Terabytes, to Petabytes.

More recently the concept of Big Data has grown out of these massive volumes of data and the need to do analytics on "unstructured" data such as text strings from social media sites or system/user logs. New technologies like Hadoop [5], Pig [6], Yarn [7], and many more are now finding a home in the lexicon of the modern IT shop. Each of these technologies carries its own price tag, and while the price point of any one technology might be decreasing, the continued need to expand into newer technologies while continuing to evolve the existing technologies, just to remain competitive, can become a nightmare to the CIO's budget. We'll now dive deeper into some of the major areas of Data Warehousing and Analytics that those CIO's need to consider.

2. DATA INTEGRATION (ETL)

We will start first with the Data Integration layer, since it is a key enabler of the initial data acquisition and responsible for much of the data movement thereafter. Of course Data Integration is much more than just moving data from point A to point B. It can also include logic for data transformations, data security, data governance, and data management.

For years, the leader in this area has been Informatica, and it continues to hold dominance in this segment. This Redwood City based company has done an amazing job of keeping pace with the evolution of architectures and technology in this space. It's one of the few companies that have been able to maintain a platform independent environment in its approach to data integration. This has been a strategic advantage for it as many of its competitors are in strategic relationships with key vendors that can hamper their ability to support other market players. A good example, of this is Ab Initio. It's an excellent, not very well known product which, because of its strategic partnership with Teradata, tends to show up primarily with just Teradata implementations.

Many other data integration players have either been acquired/absorbed by other companies, or simply created to support one specific company's products. A great example of an absorbed company is Ascential Data Stage. This was a major player in the early 2000's and one of Informatica's strongest competitors until they got purchased by IBM in 2005 and folded into their Infosphere line of products. More recently, SSIS is an example of a product that was created to support Microsoft's latest SQL Server environment. While both of these tools claim to be able to support multiple environments, they obviously are tuned to support their parent environment best.

Gartner, Inc. estimates that 2013 spending in the data integration tool market is just over $2.2 billion which is a 9.4% increase from 2012, so the interest in these tools continues to increase. In fact, if it continues at this pace it's expected to get to $3.6 billion by 2018. This high growth makes sense, because as corporate data continues to expand in both breadth and depth, the need to integrate it for usability becomes even greater.

While data integration may have started as the simple concept of extracting, transforming, and loading a data warehouse, it has grown far beyond that in the current IT shop. As the value of data has increased and its uses multiplied, data integration tools have needed to evolve to support so much more than that simple initial concept. There lies both the challenge and the opportunity – to make the data integration as simple and efficient as possible while at the same time ensuring that data governance and data security concerns are also addressed.

3. DATABASE MANAGEMENT SYSTEMS

At the heart of the data warehouse is the database management system (DBMS) [8]. This is where the data is stored and organized. Gartner defines a data warehouse as a grouping of data in which two or more separate data sources are brought together in an integrated, and time-variant strategy. Its logical design introduces flexibility across all the separate data sources without significant dependencies on the design of any existing source of data. The DBMS market was estimated at around $28.7 billion in 2013 with a 9.4% growth rate. It's expected to reach total revenue of more than $35 billion by 2016.

The critical capabilities that a data warehouse DBMS should adhere to and support include: managing large volumes of data, loading data continuously, data types other than structured, repetitive queries and advanced analytics, queries on many data types and source, operational BI queries, high system availability, multiple user skill levels (Data Scientist, Data Miner, Business Analyst, and Casual User), and standard administration/management.

Historically, the data warehouse DBMS has adhered to a relational database modeling approach with the primary warehouse conforming to a model somewhere between a third normal form and a fifth normal form. While not necessarily optimized for data access, it is by far the most efficient method for consolidating and storing the massive amounts of data dealt with in data warehousing.

The big DBMS players in the early years of data warehousing were databases like DB2, Oracle, Sybase, and SQL Server. However it wasn't long before other data warehouse and analytics optimized products, such as Teradata and HANA, quickly established themselves in the market as well. Now with the rise of new types of advanced analytics such as nontraditional, unstructured data, the DBMS market continues to evolve. Teradata has firmly established itself as the leader in the DW DBMS market and continues to work hard to maintain that spot with strategic support of the Logical Data Warehouse with technologies such as UDA, AsterData, Hadoop, and other multistructured programming constructs such as JSON or XML. Oracle has dominated the DW market for years as far as installed base, leveraging its transactional DBMS implementations to bolt on to the Data Warehouse. Gartner, Inc. estimated that it had over 42% of the Relational DBMS market in 2013. It has demonstrated itself as a solid platform for the traditional DW for years. It will be interesting to see if it can maintain that hold and continue to evolve with the market. Microsoft has also maintained a strong presence in the Relational DBMS market with a lower cost of SQL Server related products. Now with its Parallel Data Warehouse and expanded suite of related products, it's attempting to expand that influence and take market share away from its competitors. It has definite advantages for the corporate user with its easy interface to other Microsoft

products such as MS Office and Sharepoint. However it still struggles with some of the size, performance, and scalability issues that some of the other vendors have a better handle on. If we were to try to categorize data warehouse models they would fall into four primary types: Traditional, Operational, Logical, and Context Independent. The Traditional DW focuses on a collection of structured sources containing historical data. Bulk and batch loading processes are primarily used to support standard reporting and dashboarding. Its main priorities include system availability and administration, while also supporting a mixed capability query workload and breakdown based on user skills. While the Operational DW similarly uses structured data its priority is continuous loading to support applications with embedded analytics, operational data stores, and real-time data warehouse. It is mainly used for operational reporting and support of automated queries. High availability is critical and disaster recovery is very important. Support of different types of users and ad hoc queries are typically less important as the key focus is operational excellence. The Logical DW (which we'll cover in greater detail later in this paper) has a primary focus on data volume and variety with a diverse set of content data types including machine data, text, images, and video. Because these diverse content types can drive large data volumes, proper management of those large volumes is critical. It also requires meeting diverse query capabilities and skillsets including utilizing other sources beyond just the DBMS used by the data warehouse. Context Independent DW is quite unique in that it declares new data values, data form variants, and new relationships. It also supports advanced capabilities such as search and graph to assist in discovering new information models. It primarily uses free-form queries to support data science concepts including forecasting, predictive modeling, data mining, and multiple data type or multi-source queries. It tends to be used by advanced users like data scientists or business analysts because its operational requirements aare few and dependency on free-form queries across potentially multiple data types can be complex.

4. BUSINESS INTELLIGENCE

Business Intelligence and analytics is at the core of the data warehousing experience. Historically, it has been more focused on Reporting but we're now starting to see a more dramatic shift to Analysis-centric applications – so much so that some in the industry are starting to separate them into two different disciplines. In 2013, Gartner, Inc. estimated the BI platform at around $14.1 Billion and expected continued growth at around 7% annually over the following years. BI and Analytics platform capabilities include information delivery, analysis, and integration. The information delivery portion focuses on reports, dashboards, ad hoc queries, integration with MS Office (especially Excel and Access), and mobile BI. The analysis portion focuses on capabilities such as interactive visualization, search-based data discovery, geospatial and location based analytics, embedded analytics, and online analytical processing (OLAP). Finally, the integration portion focuses on supporting BI infrastructure and administration, metadata, mashup and modeling of business user data, development tools, collaboration, and support of big data concepts [9]. This space has seen the greatest amount of change over the years based on acquisitions and consolidation. Players like Business Objects, Cognos, Hyperion, Microstrategy built themselves as key application suites to support the growing BI reporting space in the 1990's. When the larger DBMS players in the market saw the opportunity of this rapidly growing market in the 2000's, they quickly began to acquire up these companies through acquisitions and mergers - SAP acquired Business Objects, Oracle acquired Hyperion, IBM acquired Cognos. However, some of them were able to hold off the acquisition push. Microstrategy, for example, has been able to maintain its independence. Then of course some companies such as Microsoft simple chose to develop their own tool suites leveraging its MS Office product line and SQL Server technology stack to develop its MSBI suite of products.

One very interesting new product worth noting which actually starts to move into the area of Advanced Analytics is IBM's Watson Analytics [10]. The product is expected to allow businesses to mine the data without the expectation of users having technical knowledge. Moreover, it would allow users to query using natural language ("Why sales are up in East Asia?".) This cloud-based tool accesses multiple datasets, correlates the data and then comes up with conclusions that are presented in a visually engaging manner to business users of every skill level. With IBM investing over $1 Billion into this initiative over the next couple of years it will be interesting to see if this becomes a game changer for this segment.

5. ADVANCED ANALYTICS

Though Advanced Analytics have been around for over twenty years, it's only recently that it has become more mainstream, gaining visibility to users beyond the small group of statisticians and scientists that had typically worked in that space. With current estimates around a $2 Billion market, Gartner, Inc. believes that Advanced Analytics will continue its rapid growth with tools becoming more user-friendly and data becoming easier to access. In fact, the recent arrival of the Data Scientist role in Corporate America is tied very closely to the discipline of Advanced Analytics. Data Science [11] is focused on improving decision making through the process of extracting key nuggets of knowledge out of complex and voluminous data. It involves various core steps ranging from basic understanding of the data, preparing the data for analysis through modeling, optimization, and simulation, and finally testing and deployment of the models into the business environment. Analytics focus on four types of knowledge acquisition: descriptive or monitoring knowledge, diagnostic or causal, predictive, and prescriptive. Descriptive or monitoring knowledge describes what is happening. This is the stronghold of traditional business intelligence platforms. Data scientists also use more advanced analytics such as machine learning to monitor and describe situations (for example clustering and decision trees). They also use anomaly detection to describe unexpected situations that could indicate danger. Diagnostic or causal knowledge yield insight into why things are happening. This has led to a new category of data discovery tools such as Tableau, Qlik, and Tibco Spotfire. Technologies and techniques for obtaining this type of knowledge included Hadoop-based data discovery, graph analysis, and case-based reasoning. Predictive knowledge helps predict what's going to happen and is a cornerstone of data science. Predictive analytics has regained much attention, driven by new use cases, as well as by new techniques such as deep learning and ensemble learning. It is

also a prerequisite for many kinds of prescriptive analytics. Prescriptive knowledge tells us how to optimize or control a given system for a certain outcome (i.e., profit maximization, market penetration, inventory turnover). There are a variety of technologies or methodologies that facilitate this supreme art of data science – for example, optimization, simulation, decision management, and predictive analytics. Advanced analytics capabilities can be grouped into the following thirteen categories: visualization & exploration, predictive analytics, data optimization, advanced descriptive analytics, simulation, further advanced analytics, data filtering & manipulation, business analytical use cases, data access, delivery integration and deployment, project and platform management, scalability & performance, and user experience. Predictive analytics and data visualization are probably the primary capabilities that have driven the expansion of advanced analytics into the corporate environment. SAS has been the undisputed leader in this area for years and is still used quite heavily by the finance and insurance industries where statistical analysis is so critical. However other companies like Oracle, IBM, and Microsoft are working hard to build out their offerings to become more competitive in these areas. Also interesting to note, some niche companies like FICO, who pioneered Credit Scoring are trying to expand their offerings beyond their niche and capture part of the market as well. One other area of Advanced Analytics worth touching on in a little more depth is Visual Analytics. Well-designed data visualization is one of the most effective ways of identifying and communicating noteworthy observations, thus easy-to-use tools, powerful tools in that space are extremely valuable. Tableau has established itself as a leader in that market. It makes a range of types of analysis accessible and easy for the ordinary business user without needing extensive training or IT assistance.

6. INFORMATION STEWARDSHIP

Gartner, Inc. estimates that 33% of Fortune 100 companies will experience an information crisis by 2017 due to their inability to adequately govern, value, and trust their enterprise information. This suggests a need for a discipline that can track and understand it. That's where Information Stewardship and concepts like Master Data Management, Data Governance, Metadata Management, and Data Quality come in. We'll briefly touch on a few of these. In general, information stewardship requires support of the following capabilities: information modeling, data quality monitoring and profiling, a business glossary, analytics dashboard, workflow, business/data rules and policy management, information life cycle or traceability, and audit/corporate memory. Master Data Management unites business and IT together with a technology based solution to ensure accuracy, semantic consistency, uniformity, stewardship, and accountability of the enterprise's formal, shared data assets. Master data can be defined as a consistent, uniform set of data and additional attributes that describe core enterprise data such as suppliers, customers, prospects, locations, organizational hierarchies and financial accounts. In other words, the primary goal of Master Data Management is to create a single "gold copy" of a data element. Two major divisions of Master Data Management, each with their own group of focused products and vendors are Product Data and Customer Data. The MDM of product data solutions market was estimated to be around $532 Million in 2013 with a growth rate of 8.7%. In 2013, the customer data solutions market for MDM was approximately $586 Million with 12.2% of projected growth. This demonstrates the value of MDM in that just these two MDM solutions combined had an investment in 2013 of over a billion dollars and growing. Key vendors in this area include IBM, Oracle, Informatica, and SAP. Some of the capabilities that MDM products support include: information quality management, business services integration & synchronization, data modeling, business process management, workflow design, scalability, availability, security, data stewardship, current technology and architecture, and information governance. Metadata Management is different than Master Data Management because it's focused on the management of the organization's information and data. Metadata describes various aspects of that data or information to help improve its usability life cycle. It supports capabilities such as: metadata repositories, business glossary, data lineage, impact analysis, rule management, semantic frameworks, and metadata ingestion and translation. As you can surmise from this list of capabilities, Metadata Management represents the tactical deliverables of an information management program. Data Quality, on the other hand, is a focuses on the business process and the accuracy of the data used by those processes. It goes beyond mere technology and focuses on concepts such as organizational roles & structures; monitoring, measuring, reporting, & remediating data issues; and and higher concepts like Policy creation and information governance activities. Its primary capabilities support: data profiling, measuring data quality, parsing, data cleansing, standardization, matching, monitoring, issue resolution, workflow, and data enrichment. Gartner, Inc. estimates that the data quality market was $1.13 Billion in 2013 and that growth in this market will accelerate over the next few years at almost 16%. It's one of the fastest growing markets in the enterprise software sector. Approximately 50% of which is controlled by several large and well established vendors including SAP, Pitney Bowes, Informatica, and Trillium. One other interesting point to note, is that Party (customer) data continues to be the primary data quality focus for 86% of those polled.

7. NEW DATA WAREHOUSE CONCEPTS

Big Data has definitely been a hot topic the past couple of years. It can be defined with the three "V's" of variety, velocity, and/or volume, which in turn focus on cost effective information assets, information processing innovation, and enhanced insight into process automation and good decision making. It's typically associated with large unstructured data such as "click stream", social media, sensor, and text based data sets, and is probably the biggest consumer of cloud based computing. It will be of critical importance that future data warehouses efficiently integrate big data into their environments and that's where the Logical Data Warehouse (LDW) comes in. Over the past five years, we have begun to see the rise of the Logical Data Warehouse. The LDW principally sees the data warehouse as a data management platform integrated for data asset consolidation as well as time variant support in its use. More importantly it is not specifically seen as a centralized physical repository of data. Repositories specific to the data warehouse are only one part of the data warehouse and a small part of the overall SLA needed for data management. Data warehouses should avoid the limitation that occurs when using a single platform for

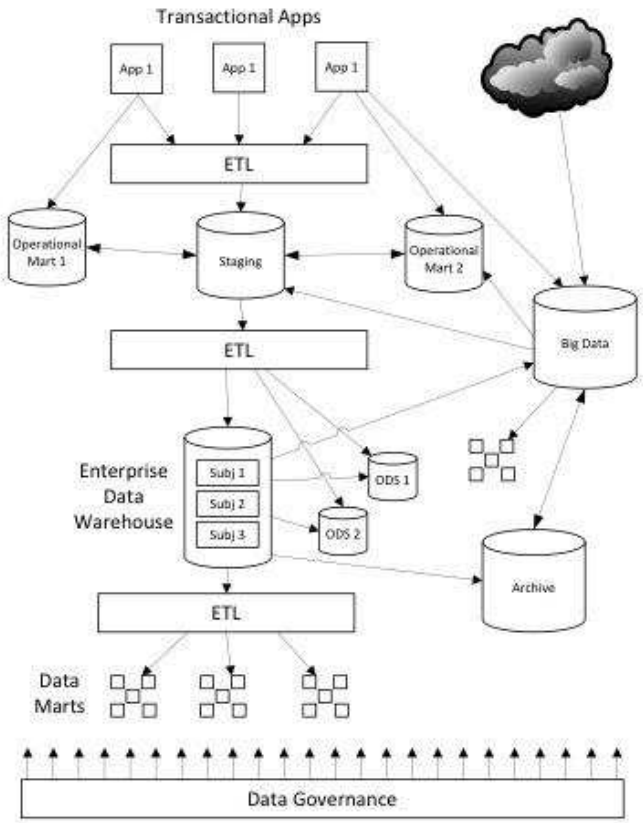

Figure 1: The Logical Data Warehouse

data management. The LDW is typically comprised of seven major components: SLA management, repository management, distributed processes, data virtualization, statistical audits, metadata management, performance evaluation services, ontology & taxonomy resolution. You can see an example of how it all fits together in Figure 1. Data comes into the picture in a variety of ways and depending on the analytical SLA requirements it can get staged for consumption at any number of levels. Undergirding it all is data governance tools and processes, because without that you could quickly lose control of the entire data management structure.

Two other technologies that are worth noting are Hybrid Transaction Analytical Processing (HTAP) and In Memory Computing (IMC). Both technologies offer key support of the Logical Data Warehouse concept. We won't go into these technologies in great detail, but essentially they combine transactional processing and analytical processing into the same database enabling applications to analyze data while it's still being created or being updated through normal transaction-processing functions. In memory computing will be a key enabler of HTAP and is founded on the principle that all application data needed for processing is located within their computing environment's main memory, which in turn is enabled by the rapidly dropping cost of memory technology and by specific application infrastructure software and cloud services. In other words, the computer's main memory can hold, at a reasonable cost, large datasets (up to multiterabytes, either on a single server or through the clustering of multiple servers) for sharing across several, possibly distributed, applications.

8. CONCLUSION

Even though data warehouse practices have remained fairly constant over the past thirty years, with the current rate of technological change and growing analytical demands, Gartner predicts that by 2019 traditional data warehouse practices will no longer be relevant. Data warehouses will have to go far beyond reporting and traditional style business intelligence, and tackle challenging requirements such as integrated information support with differing analytic use cases, incompatible service-level expectations, operational application embedded analytics & data provisioning, and hybrid transaction & analytics processing. The ability to make the shift in order to stay competitive will be key for all those involved including the architecture, the product vendors, and data warehouse and analytic professionals that pull it all together and keep it running. The rapid advances we continue to see in data storage and processing coupled with new types of data analytics, make this an extremely exciting time to be part of this world. As we strive to fulfill our thirst for knowledge, we expect these new technologies and data architectures to have a similar impact on the world of information management that the industrial revolution had on the manufacturing world over a century ago. Perhaps we should start referring to it as the Information Revolution.

9. REFERENCES

[1] C. Beath, I. Becerra-Fernandez, J. Ross, and J. Short, "Finding Value in the Information Explosion", MIT Sloan Management Review, 2012.

[2] W. H. Inmon, "Building the Data Warehouse", Wiley, 2005.

[3] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, "The Data Warehouse Lifecycle Toolkit", Wiley, 2008.

[4] E. Thomsen, "OLAP Solutions: Building Multidimensional Information Systems", Wiley, 2002.

[5] T. White, "Hadoop: The Definitive Guide", O'Reilly, 2015.

[6] A. Gates, "Programming Pig", O'Reilly, 2011.

[7] A. Murthy, V. Vavilapalli, D. Eadline, J. Niemiec, and J. Markham , "Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2", Addison-Wesley Data & Analytics, 2014

[8] C. Coronel and S. Morris, "Database Systems: Design, Implementation, & Management", Cengage Learning, 2016

[9] S. Williams, "Business Intelligence Strategy and Big Data Analytics: A General Management Perspective", Morgan Kaufmann, 2016.

[10] J. Miller, "Learning IBM Watson Analytics", Packt Publishing, 2016

[11] F. Provost and T. Fawcett, "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking", O'Reilly, 2013.

[12] H. Zhang, G. Chen, B. C. Ooi, K. Tan, and M. Zhang "In-Memory Big Data Management and Processing: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 7, July 2015.