An Eye on Enterprise Data Warehousing Tools and Best Practices

Ms. Rupali Chikhale, Dr. Sheel Ghule
G. H. Raisoni Institute of Information Technology, Nagpur

Abstract – Most of the companies implementing a strategy for multiple business intelligence (BI) data warehouses to provide significantly more powerful analytics capabilities to business groups. By providing an array of BI platforms, we are helping a broader range of data faster, deeper, and more cost-effectively. This expanded architecture enables our business groups to solve more high-value business problems, achieve greater operational efficiencies, and improve their competitive performance in global markets. The business environment has become instantaneous. Users need very rapid access to more insights and they cannot afford to wait—else they lose a competitive edge. For IT organizations, this means delivery of relevant, timely insights faster than ever before. Objective of this paper is to provide clear starting points that you can leverage to redesign your architecture. Alternative delivery models—including software, appliances, cloud options— help drive faster time to value. As a pioneer in infusing cognitive capabilities into its data warehouse strategy, organizations can use it as a starting point for improving their DW/BI solution.

Keywords: DW, EDW, BI, ETL, OLAP.

I. INTRODUCTION

In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

II. DATA WAREHOUSES VERSUS OPERATIONAL SYSTEMS

Operational systems are optimized for preservation of data integrity and speed of recording of business transactions through use of database normalization and an entity-relationship model. Operational system designers generally follow the Codd rules of database normalization in order to ensure data integrity. Codd defined five increasingly stringent rules of normalization. Fully normalized database designs (that is, those satisfying all five Codd rules) often result in information from a business transaction being stored in dozens to hundreds of tables. Relational databases are efficient at managing the relationships between these tables. The databases have very fast insert/update performance because only a small amount of data in those tables is affected each time a transaction is processed. Finally, in order to improve performance, older data are usually periodically purged from operational systems.

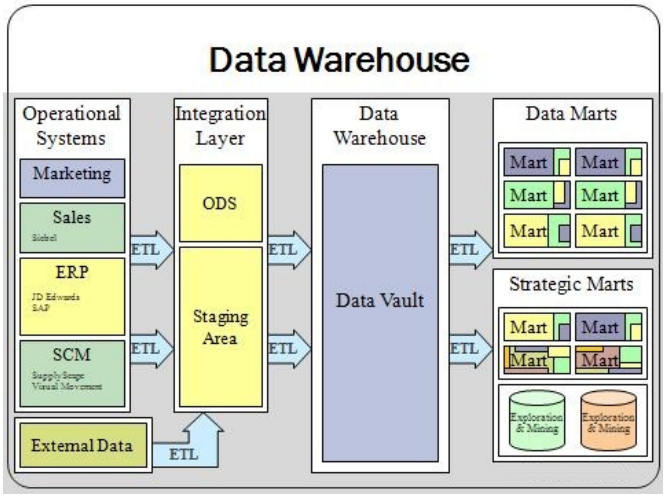

Fig. 1 DW Overview

Data warehouses are optimized for analytic access patterns. Analytic access patterns generally involve selecting specific fields and rarely if ever 'select *' as is more common in operational databases. Because of these differences in access patterns, operational databases (loosely, OLTP) benefit from the use of a row-oriented DBMS whereas analytics databases (loosely, OLAP) benefit from the use of a column-oriented DBMS. Unlike operational systems which maintain a snapshot of the business, data warehouses generally maintain an infinite history which is implemented through ETL processes that periodically migrate data from the operational systems over to the data warehouse.

III. SCOPE OF DW TOOLS

Today, data warehousing tools typically need to perform the following major operations:

- Batch and near real-time loads to integrate data from multiple resources (internal and external)
- Basic reporting with no drill-down/ drill-across
- Online analytical processing (OLAP)
- Predictive analytics
- Operational business intelligence

With time, data warehousing tools are required to meet a set of new challenges. Some of these are discussed in detail here.

Big data

Some of the big challenges facing data warehousing tools today are exploding data volumes, new emerging types of data, more real-time latencies, the lack of agility in delivering data to the business, and more mission critical use of data warehouses in operations and decision making. "The ratio of growth of data vis-à-vis that of storage capacity is about 2:1. Analyzing such rapidly growing data volumes and extracting business value from them will be a key challenge to data warehousing tools," says Suganthi Shivkumar, managing director - Asia South at Informatica. To meet the challenges of real-time operations and big data, Informatica has released native connectivity with Hadoop in June 2011. This enables customers to deliver all types of data at varying latencies. IBM, another large vendor, has employed its core technologies InfoSphere, BigInsights, and InfoSphere Streams as its platform for big data.

ETL

For enterprise data warehousing tools, the Extract, Transform and Load (ETL) operations from the information sources represent another challenge. Doing so in real-time is tough further. Most ETL tools operate in a batch mode on the assumption that information will be available on a certain schedule. When operating in real-time mode, it's a lot more challenging as the ETL operations need to happen simultaneously when the transactional systems are experiencing peak loads. "As OLAP and query tools are designed to operate on unchanging data, their operations on real time data can lead to inconsistent and confusing results," says Sheshagiri Anegondi, vice-president – technology business at Oracle India.

IV. COMPLEX QUERIES

At present, many companies are using traditional, proprietary data warehousing tools that aren't designed to handle complex analytic queries against billions of rows of data. To answer even simple questions typically requires time-consuming retooling, creating indexes, partitioning the data and re-indexing the database.

Commenting on how the data warehousing tools will look like in the near future, Sanjay Raj, director of BI/DW practice at Syntel, says, "Doing all that the business wants will need a design that is modular, scalable and can sustain the performance required."

V. NEW TRENDS

1. Cloud computing

Enterprise data warehousing tools with cloud/ SaaS model have come into the enterprise. "Over the next two to three years, these will gain greater enterprise adoption as a complement or outright replacement for appliance- and software-based data warehousing tools," says Kobielus of Forrester. Adds Syntel's Raj, "Public clouds will take some time to mature because of security concerns. Private cloud is a reality and many organizations in the services and BFSI sectors are aggressively pushing it."

2. Transformational technologies

The technologies like in-database analytics and transaction processing can transform the role of enterprise data warehousing tools. The current best-of-breed platforms support these application integration scenarios through features and interfaces such as MapReduce, in-database function pushdown, embedded statistical algorithm libraries, predictive modeling integration, decision automation, and mixed workload management.

3. Social media and unstructured data

Social media drove unstructured data and real-time architectures into the enterprise data warehouse. A key application is social media analytic dashboards to monitor customer awareness, sentiment, and propensities in real-time. To address these requirements, and the convergence of in-database data mining and text analytics, next-generation data warehousing tools incorporate unstructured sources, hybrid

storage architectures, in-memory execution, distributed cache, complex event processing, solid-state drives, geospatial data sets, and rich metadata.

4. No SQL

Most commercial applications and solutions use a relational database under the hood for metadata/content store. Non-relational distributed NoSQL databases would have to go through a long cycle of time test to convince solution developers and business users in order to create a market place. Customers are highly sensitive to the application and database support availability of the solution; however, NoSQL databases being an open source offering, will have to address and overcome support challenges before it can be part of any critical business system.

5. In-memory technology

In-memory databases catering to sub-second response requirements do not share any common business space with data warehousing tools. Informatica is working closely and innovating with other data warehouse vendors in these areas; for example, Greenplum, Teradata/Aster Data, HP/Vertica, etc. Oracle acquired TimesTen in 2005 and today it forms the cornerstone of Oracle's offerings in in-memory databases. However, data warehousing tools are marching towards in-memory paradigm with the advent of solid state drives in order to achieve higher performance.

6. Open Source

Open source isn't new, of course. When the Internet took flight in the mid-1990s, Linux sparked a free software movement that today supports everything from operating systems to application servers to middleware and databases. So, why is open source a particularly smart strategy for data warehousing tools?

As of today, many features that are part of proprietary products are not available in open source analytical databases. But, some of the features like partitioning, bitmapped indexes, materialized views, parallel loading and query processing, support for SQL Windowing functions that were missing earlier are now being made available making the sun of open source data warehousing tools shine brighter.

Some best practices:

Best practice 1: Ensure support and sponsorship from the CEO's desk

As a data warehousing best practice, while considering investments, ensure executive buy-in. The sponsor of the data warehousing project plays a key role and it's desirable that the CEO undertakes it. If the IT department propagates the data warehousing initiative, it will be viewed as an IT project and business users may lose interest in it.

Consider data warehousing as an organization-wide project and not a departmental one. The initiative may start on a departmental level for trial and error iteration, but gradually, it has to span across the enterprise.

Best practice 2: Develop a holistic RoI model for DW investments

The data warehousing project planning should identify the key performance indicators (KPIs) to measure aspects like revenue increase and operational ease. These factors should be quantified to the best of one's ability.

A major bottleneck for data warehousing is that these projects are looked upon as additional costs. They should rather be viewed as enablers to improve the top line, bottom line, and operational efficiency.

Best practice 3: Consolidate historical data by beginning data warehousing early

As a data warehousing best practice, begin investing as soon as the organization implements complex operational systems like enterprise resource planning or customer relationship management. These systems are fertile grounds for data generation. Consolidating the data from these sources will immediately aid in establishing and maintaining good data quality. The consolidated data will fuel faster and more accurate business intelligence (BI).

Best practice 4: Know the domain

The DW will reflect how different businesses interact and interface with each other to meet enterprise-wide objectives. To illustrate, if a bank sets out to develop a DW, banking-specific domain expertise will be needed to customize the solution.

Best practice 5: Know the source systems well

Data warehousing is about integration. Another consideration is to find if the functional and technical nuances of the source

systems interact collectively or work individually. For instance, a bank will have a core banking system, a credit card system, a PRA (purchase and resale agreement) system, and others. The method in which these systems generate data varies. Having knowledge to incorporate these variations in data warehousing is important.

Best practice 6: Plan for flexibility and extendibility of data warehouse solution

A key data warehousing best practice is to ensure that the data model is flexible. The model should be able to extract data from additional source systems. The data model should not just address the current acute needs, but also suffice when these demands grow, without redesigning the entire DW. Flexibility and extendibility are important features that you can equip your DW with. Many a times, scalability is considered only from a performance point of view. Extendibility would entail how easily can more subject areas be added and can columns be added to store excess data. Best practice 7: Make metadata and impact analysis intuitive, integrated

Metadata, which is a manifestation of the three levels—data model, extract, transform load (ETL), and BI reporting—of a data warehouse, needs to be intuitive. The names of the columns and tables should not be abstract, which nobody would automatically relate to. The informal way of data warehousing is writing scripts and not worrying about metadata. However, the formal way would entail having the technical infrastructure in line with the business strategies and managing the data as well as the changes in the DW. A common observation is that adding of a new source system impacts current ETL modes. A formal data warehouse with well-built metadata layer will show the jobs that are impacted. However, when working in the script mode, the areas of impact will have to be checked manually and this could be time consuming, cumbersome, and costly. Best practice 8: Maintain balance and control through data governance and auditing

Audit the data regularly as a part of the data warehousing quality and control process. This will ensure that the data from the source systems and reports match. A data warehousing best practice is to be proactive with scripts written to monitor the usage of data models and reports. Automated notifications will appear in case of mismatches between data from the source and DW. Reports about these gaps should be generated on a daily basis. To maintain data quality, a data champion could be employed.

Best practice 9: Plan and provision for ETL for future growth In a DW, the ETL infrastructure would contain an ETL tool, servers, and database. If you choose the ETL tool that comes as part of a BI package, you may face flexibility issues. Such tools may not accommodate the source systems and additional data needs. As a data warehousing best practice, take the effort to evaluate and buy an appropriate ETL tool.

Best practice 10: Ensure that the project team has sufficient knowledge of BI

Generally, data warehousing projects are undertaken when organizations plan to deploy business intelligence. From this context, it becomes necessary to evaluate how business is conducted. Such evaluation will form as the foundation of any conversation on developing a data warehouse. The team that works on the DW development, should, therefore, have a fair understanding of business intelligence.

VI. CONCLUSION

The main conclusion from the study is that, it help organizations improve their DW solutions, there is no —silver bullet for a successful development of DW/BI solutions. Here it provides a quick way for organizations to assess their DW/BI maturity and compare themselves in an objective way against others in the same industry or across industries.

Providing multiple BI data warehouses greatly expands the ability of business groups across Intel to mine the enormous amounts of raw and unstructured data. Matching the use case with the most appropriate BI platform is enabling us to achieve substantial cost savings. The previous approach of relying on a single, centralized data warehouse became both costly and limited for our expanding BI needs, so revising our BI strategy to accommodate multiple data warehouses can significantly enrich the decision making process across the company and enhance business performance.

In the several business use cases where this strategy has been employed, BI solutions have been generated avoiding the use of the more-costly EDW platform. We will continue to analyze the financial benefits, basing them on the overall benefits derived from these use cases.

REFERENCES

[1] Evolving Application Domains of Data Warehousing and Mining: Trends and Solutions Furtado, Pedro Nuno San-Banto IGI Global, 30-Sep-2009 - Computers - 364 pages

[2] http://www.prithviinc.com/PrithviBIDW_CaseStudy.pdf.

[3] http://www.slideshare.net/huongcokho/data-mining-concepts

[4] Fayyad U, et al 1996: From Data Mining to Knowledge Discovery: An overview. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

[5] Gibert K, A. García-Rudolph, G. Rodríguez-Silva 2008: The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. Acta Informatica Medica 16(4) 178-182

[6] Spate J, K. Gibert, M. Sànchez-Marrè, E. Frank, J. Comas, I. Athanasiadis, R. Letcher 2006. Data Mining as a tool for environmental scientist. In procs 1srts iEMSs Workshop DM-TES 2006 Third Biennal Meeting: "Summit on Environmental Modelling and Software". Burlington, VE USA.

[7] Vazirigiannis M, M. Halkidi, D. Gunopulos 2003: Uncertainty handling and quality assessment in data mining. Springer-Verlag.

[8] https://iaonline.theiia.org/data-mining-101-tools-and-techniques

[9] http://media15.connectedsocialmedia.com/intel/02/12555/Using_Multiple_Data_Warehouse_Strategy_Improve_BI_Analytics.pdf

[10] http://www.dataminingtechniques.net

[11] "An architecture for a business and information system". IBM Systems Journal.

[12] http://searchbusinessintelligence.techtarget.in/tip/Data-warehousing-best-practices-Part-I

[13] Jump up ^ Inmon, Bill (1992). Building the Data Warehouse. Wiley. ISBN 0-471-56960-7.