Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 2, February 2014, pg.508 – 516

RESEARCH ARTICLE

Data Integration Models for Operational Data Warehousing

G. Swetha[1], D. Karunanithi[2], K. Aiswarya Lakshmi[3]

[1]M. Tech. Scholar, Department of Information Technology
Hindustan University, Chennai, Tamilnadu, India
swethamohan.23@gmail.com

[2]Assistant Professor, Department of Information and Technology
Hindustan University, Chennai, Tamilnadu, India
karunanithid@gmail.com

[3]M. Tech. Scholar, Department of Information Technology
Hindustan University, Chennai, Tamilnadu, India
lakshmi.karakkal@gmail.com

ABSTRACT

Data warehouses have evolved to support more than just strategic reporting, analytics and daily forecasting. Organizations are investing significant resources to integrate valuable information contained in their data warehouse into their day-to-day operations. Incorporating business intelligence into decision making enables these organizations to optimize business performance throughout the day. However, to achieve these efficiencies, data must be provided in real time environment. There are many data integration technologies that serve the data acquisition needs of a data warehouse in organizations, and the demand for low-latency data is causing IT organizations to evaluate a wide range of approaches: intraday batch Extract, Transform, and Load (ETL) processes as well as real-time Change Data Capture (CDC) techniques.

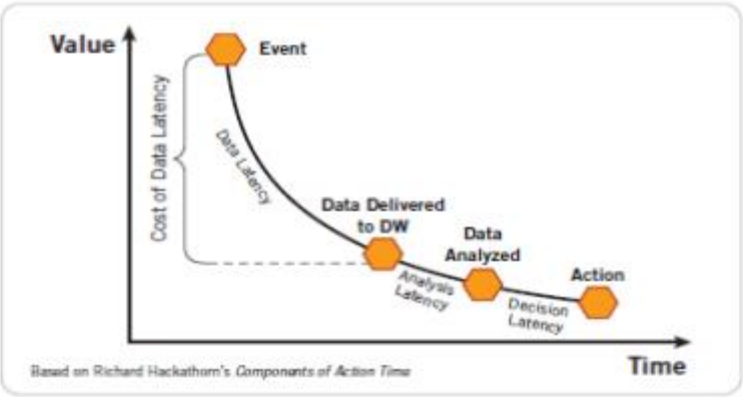
1. INTRODUCTION

Business time is increasingly moving toward real time. As organizations look to grow their competitive advantage, they are trying to uncover opportunities to capture and respond to business events faster and more rigorously than ever. Today, the majority of competitive advantage comes from the effective use of IT. Therefore, from that standpoint, the key to achieving faster and accurate Business Intelligence (BI) is a robust enterprise data warehouse combined with an enterprise analytics framework.

Why Real-Time Data for the Data Warehouse?

Across the enterprise, [2]each fact of the business gathers data through an assortment of activities, and many organizations now deliver this data to a central data warehouse—where the data is captured, aggregated, analyzed, and leveraged to improve decision making. The quality of these decisions depends not only on the sophistication level of the analytics applications that run on the data warehouse, but also on the underlying data. Data has to be complete, accurate, and trusted. For that reason, it has to be timely: timely data ensures better-informed decisions.

To approach real time, the duration between the event and its consequent action needs to be minimized. As outlined in Figure 1, the initial data acquisition and delivery to the warehouse introduces the majority of the latency.

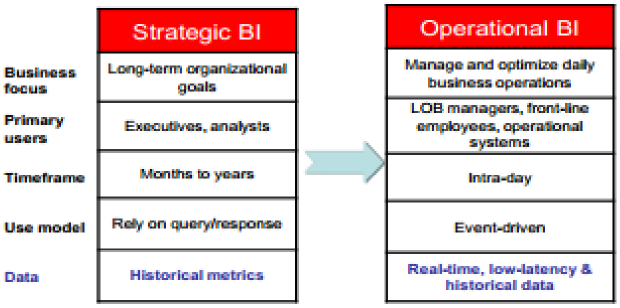


Based on Richard Hackathorn's *Components of Action Time*

Figure 1. The longer it takes to capture and process data, the lower the value of the information.

The Data Warehouse Evolution to Operational Decision Support for Front-Line Users Traditional data warehouses have focused on support for strategic BI—[1]a resource for the small group of analysts and decision makers engaged in strategic planning that affects time horizons of months or years. Today, more and more companies maximize the value and competitive advantage of their data warehouse by using it in an operational role, adding mission-critical decision support to their workload.

This new capability is referred to as operational BI



	Strategic BI	Operational BI
Business focus	Long-term organizational goals	Manage and optimize daily business operations
Primary users	Executives, analysts	LOB managers, front-line employees, operational systems
Timeframe	Months to years	Intra-day
Use model	Rely on query/response	Event-driven
Data	Historical metrics	Real-time, low-latency & historical data

Based on Claudia Imhoff's "Differences between strategic, tactical and operational BI".

Figure 2. Data warehouses have moved beyond strategic planning into the realm of operational resources.

Long-term strategic decision making could be based on historical metrics derived from day-old or week-old data. In today's competitive business world, companies need to see ROI from their data warehouse and BI investments, not just in strategic planning but also in operational decision making.

Particularly, front-line employees can provide more-responsive service and create efficiencies in their business functions if they have the most up-to-date information possible. By combining historical data with ongoing operational data, operational data warehouses enable a much-larger population of business users to make more-informed, proactive decisions.

The [3] enterprise data warehouse helps personnel make excellent “small decisions” that collectively enhance competitive advantage in line with business strategy. For example:

- Can I expedite this package based on the criteria I see here?
- Can I extend a special offer or up-sell to this customer at this moment?
- Can I change our current marketing campaign based on this morning’s result

There are various business areas real-time information can be leveraged to gain competitive advantage.

Below are the top 5 areas where you can expect a difference in customer experience, or in operational efficiency, or in both.

1.1. CUSTOMER SERVICE

By using up-to-the-minute data businesses gain a complete, up-to-date view of the customer, including customers’ most recent activity on the company’s website or the latest status of a service ticket they submitted. When the customer contacts the company, the service team has the information to handle their request fast and effectively.

In addition, nowadays companies need to provide up-to-date information to their customer via customer portals. By updating customer portals with real-time data from back-end systems, companies can offer accurate information to their customers on the Web. This, in return, decreases the number of calls customers make to access, confirm, or correct information about their account, resulting in operational cost savings for the company.

1.2. TARGETED MARKETING

Up-to-date information about customers also enables personalized, targeted campaigns when the customer is ready to engage with you. Several industry leaders use real-time information to offer personalized promotions to their customers considering the issue they are contacting the company for.

The result they see is higher acceptance of promotional offers, since the offer is relevant to customers' current issue and the customer is already in contact with the company.

1.3. CAMPAIGN OPTIMIZATION

The sooner the marketing team sees the results of their ongoing campaign, the sooner they can adjust their promotion and increase their return on the campaign effort.[5] This capability enables to optimize marketing budget and achieve campaign targets faster. For example, in the retail industry companies use real-time data integration to gather store data from regional locations back to headquarters. They use the data for timely comparisons of stores' results and improve the way they execute their promotions.

1.4. FRAUD DETECTION

Similar to above, the earlier a risky transaction is identified; the earlier it can be stopped; preventing further loss. For this reason, many leading organizations in financial services and telecommunication industries maximize ROI on their fraud detection systems by bringing the data from transactional systems in lowest latency possible.

1.5. WORKFORCE OPTIMIZATION

For service-based businesses, optimizing human resources and where they are performing their job can not only improve customer experience but also minimize labor cost. For example, in field service operations employees' schedules can easily change throughout the day due to external circumstances, such as traffic, or customer demand. Real-time data integration allows the resource planning applications to factor in the dynamically changing events to provide the most efficient schedule and workforce distribution to service the customers.

Business intelligence for operational execution drives tangible benefits many other areas including in operations facilitates automation, which improves efficiency.

2. DATA INTEGRATION APPROACHES FOR OPERATIONAL DATA WAREHOUSING

There are numerous technologies that serve data acquisition needs. One of the biggest differentiators among these solutions is the speed of data capture and delivery, as well as impact on the source systems. Only a few offer real-time data delivery with low system impact and no reliance on batch windows. Choosing the right solution requires a comprehensive understanding of organizational data requirements, including

- Data volume (size of data and number of updates)
- Date movement frequency
- Data integrity
- Transformation requirements
- Outage windows required/impact on business continuity (batch windows)

To clarify a common misconception, some data acquisition technologies often refer to “right-time” BI.

Right time refers to the needs of the end users in accessing intelligence and can be different across different use cases. The need for data latency also changes over time. Which data latency users need today may change in the next year or few months, depending on the projects and new business initiatives. For operational data warehousing, the underlying technology infrastructure should deliver real-time data integration capabilities and let the business user choose the right time to access the data.

3. TRADITIONAL DATA INTEGRATION APPROACHES

Traditional data acquisition approaches include scripting, ETL, EAI, and real-time CDC. Scripts and ETL are batch oriented in data delivery, whereas EAI and real-time, log-based CDC support continuous data capture.

3.1 SCRIPTS

Scripts are flexible and economical to develop, and almost every operating system can invoke scripts from their built-in scheduling data. However, scripts pose many challenges, such as being a drain on developer resource time and effort, as well as administrative challenges, such as manageability, documentation, and service-level agreement compliance.

4. EXTRACT, TRANSFORM, AND LOAD

ETL can be an ideal solution for the bulk movement of large volumes of data. Packaged ETL products also offer advanced transformation capabilities. As for data acquisition, ETL tasks are executed intermittently—typically during all maintenance windows when the data sources are available, to ensure that data sources don’t change during data acquisition and lead to inconsistencies across online transaction processing (OLTP) systems and the data warehouse.

To support this configuration, for the most part ETL tools must store additional data in source tables, such as time stamps, to identify changed data since the last query. Most databases were not designed

To decrease data latency, some ETL products can perform or be customized for CDC capabilities.

To support this configuration, for the most part ETL tools must store additional data in source tables, such as time stamps, to identify changed data since the last query. Most databases were not designed reporting. The real-time CDC solutions that capture the changed data from the database transaction logs do not impact the performance on the source systems, unlike offerings that use database triggers or table scanning.

Data Transformations -From where we got it?

As data warehouses evolve and become more operational with the benefit of real-time data feeds, the requirements for transforming the data have also changed. As previously described, in traditional data warehousing, data acquisition tends to be batch oriented. Data moves between relational and multidimensional structures, and typically most of the transformations are handled on the chosen ETL engine.

As the data warehouse approaches real time, transformations tend to take place in the data warehouse. This is often called an ELT approach: extract, load, and then transform. The data warehouse stages and transforms the data to reduce data and analysis latency. This eliminates the need to aggregate changed data on a centralized server and removes an intermediate step from the overall data flow, as well as the associated costs of acquiring and managing the dedicated ETL server.

A major requirement for operational data warehouses that receive real-time data feeds is to handle both loading and querying workloads simultaneously. Enterprise data warehouses are increasingly being designed to support these mixed workloads so that the benefits of real-time data feeds can be fully realized. Leaders in data warehousing solutions, such as Oracle Exadata, support mixed workloads, enabling continuous data loading, dashboard updating, and prebuilt reporting with timely data.

5. CONCLUSION

Succeeding in today's competitive business environment requires good decisions, not just at the top level of the organization. Operational data warehousing allows all users in the organization to access and respond to information in a timely manner. Establishing and maintaining this real-time data warehouse requires a continuous low-latency data capture and delivery infrastructure.

- Real-time data for enabling more-advanced, agile BI
- Low-impact, high-performance data integration by reading database transaction logs
- Zero requirement for batch windows or using a middle-tier server
- Integration with Oracle Data Integrator EE 11g for high-performance ELT architecture
- Support for large data volumes and heterogeneity
- Ability to augment existing ETL solutions with real-time, low-impact data acquisition
- Exceptional flexibility, easy implementation, and maintenance
- Robust data recovery after outages in organisations
- Ability to move read-consistent data with referential integrity

Organizations that leverage the most up-to-date BI in their day-to-day operations have seen significant improvements in operational quality, productivity, and customer service

References

[1] R.A. Coyne and R.W. Watson,“The Parallel I/O Architecture of the High-Performance Storage System (HPSS),” Proc. IEEE 14th Symp. Mass Storage Systems (MSS), 1995

[2] Nat’l Center for Biotech Info, http://www.ncbi.nlm.nih.gov/, 2013.

[3] Pbs Pro Technical Overview: Scheduling and File Staging,

https:// secure.altair.com/sched_staging.html, 2008.

[4] Cluster Resources Inc., http://www.clusterresources.com/, 2008.

[5] T. Kosar and M. Livny, “Stork: Making Data Placement a First Class Citizen in the Grid,” Proc. 24th Int’l Conf. Distributed Computing Systems (ICDCS), 2004.