Taming The Data Beast

INTEGRATING SYSTEMS AND APPLICATIONS FOR A BETTER BUSINESS

Prepared by TechRepublic exclusively for Dell
CONTENTS:

Data, Data Everywhere .................................................................3
Making Information ........................................................................4
Data: Your Most Critical Asset .........................................................5
Paradigm Shift: From ETL to CDC-TL ................................................5
Beginning the Process - Know and Love Your Data ...............................7
Summary ..........................................................................................8
About Dell .......................................................................................9

About this resource guide: This guide has been prepared by CBS Interactive on behalf of Dell. Dell has specified topic, title and key themes of this guide and may have contributed to and exercised editorial control over the content. This guide may only be quoted and reproduced by Dell in its entirety.
“It is a capital mistake to theorize before one has data.”  
—Sherlock Holmes, *A Study in Scarlett (Sir Arthur Conan Doyle)*

“Data are becoming the new raw material of business.”  
—Craig Mundie, *head of research and strategy, Microsoft*

**DATA, DATA EVERYWHERE**

In 1970, Edgar Codd published the paper, “A Relational Model of Data for Large Shared Data Banks,” essentially inventing the relational database and setting the stage for the likes of Oracle and IBM to create the multi-billion dollar Relational Database Management System (RDBMS) industry. Shortly after, Larry Ellison founded Oracle, IBM revolutionized the retail industry and transaction processing by introducing UPCs and barcode scanners, and the commoditization of personal computer hardware made point-of-sale and other types of end-user data collection the norm.

Fast forward to today: databases utilize the same principles that underlie massive data warehouses.

Although IBM and Oracle remain among the largest technology companies in the world, the growth of the Internet has surfaced database-driven user experiences delivered by hundreds of thousands of companies from around the globe. At the same time, e-commerce, advances in real-time data collection and monitoring, and, more recently, the anywhere/anytime mindset of mobile computing has organizations regularly measuring data in terabytes and petabytes instead of kilobytes.

The data collected on customers, sales trends, market projections, and web impressions is both the greatest equalizer and the greatest differentiator of our time. Businesses of all sizes need to aggregate, transform and use data from:

- Sales transactions
- Online transaction processing (OLTP) systems (in some cases, OLTP represents the end product of stored, analyzable data; in other cases it is simply an intermediate aggregation of other data)
- Web traffic
- Surveys and market research
- Advertising campaigns
- Customer demographics and contacts
- Human resources
- Finance
- Manufacturing
- Supply chain data and enterprise resource information
- Sensor data and equipment logs
- Social data from various online networks and communities
Boutique retailers have access to the same CRM software (specialized databases) as JCPenney, and a small organic coffee roaster can leverage the same web analytics tools and data-driven ad networks as Starbucks.

MAKING INFORMATION

The real trick, of course, is taking all of this data and turning it into actionable information. Most businesses have become fairly adept at collecting lots of data, but finding the most efficient and effective means of aggregating the data and ultimately using it to inform business decisions has become vastly more difficult. The Business Intelligence (BI) industry is devoted to helping organizations make decisions on-the-fly based on wide-ranging data collected from multiple sources and then subjected to various business rules.

BI has historically been applied to everything from the early identification of potential high school dropouts to automated changes in manufacturing and ordering.

Even for businesses whose data collection processes and/or scale don’t support automated (and expensive) BI systems, the science (and many would say the art) of data analytics is critical to objective, strategic, data-driven decision-making. Perhaps one of the best examples of converting raw data to information is Google’s Flu Tracker.

Based on search queries related to influenza, Google is able to determine peak flu seasons across the country and pinpoint local epidemics with greater speed and accuracy than the Centers for Disease Control and Prevention. Google has applied this same approach to tracking the movement of illegal arms and the activities of drug cartels.

**Historical estimates**

*United States Flu Activity*

*Influenza estimate*  

*Google Flu Trends estimate*  

*United States data*

(Image courtesy of Google)
DATA: YOUR MOST CRITICAL ASSET

A recent summer internship program with IBM allowed college and graduate students to design projects and take deep dives into the capabilities of Watson, the company’s supercomputer made famous by competing on Jeopardy. Half of the participating students were, not surprisingly, studying computer science and related fields; however, the other half were studying business, marketing, healthcare, and other fields outside of the expected computing arenas. They were interested in supercomputing applications for dealing with increasingly challenging data problems facing the business world.

There is no more critical asset that a company or organization possesses than its data. So, how would Watson fit into all of this?

Data comes from so many sources and is often so difficult to reconcile, especially as the sheer quantity of data increases, that it will eventually take a supercomputer capable of human-like synthesis (albeit at much higher speeds) to make sense of it all.

Watson aside, a multitude of tools already exist for ensuring the safety, integrity, usability and viability of data stores. Even small organizations can synthesize multiple data points to drive strategic planning. These tools generally take the form of Decision Support Systems (DSS) and are critical to timely, effective decision-making based on data from multiple sources.

PARADIGM SHIFT: FROM ETL TO CDC-TL

One of the most challenging aspects of “taming the data beast” is dealing with disparate data sources that don’t align naturally to form cohesive data sets. A business can have the most brilliantly normalized, well-conceived data warehouse with tables and fields that beg for analysis and BI applications. But if the CRM application delivers nightly data dumps that amount to a square peg in the warehouse’s round holes, the whole business of synthesizing information from the data store goes out the window.

The best solution is a tight integration of an organization’s applications with its data warehouse. However, because few data-generating applications are designed by the same vendor with the express purpose of long-term compatibility, this is far easier said than done. In the same way, because systems may be custom built, incorporate elements from different verticals, contain legacy structures and applications, and may not be intended to talk to one another, there are rarely standardized hooks that facilitate back-end integration.

Thus, any solid data warehousing, mining, and analytics strategy will include a laser focus on Extract, Transform, Load (ETL) architectures and management of the potentially millions of transactions that go into it. ETL refers to the process by which data from any number of sources are processed to conform to the structure of the data warehouse in which they are ultimately integrated and will reside for analysis and mining.
ETL processes then are essentially brokers between applications and data warehouses (or so-called datamarts, or operational data stores if the end product is more specific or temporary in nature). ETL itself, though, must not only be carefully designed and fully vetted but also tracked and audited. This is, after all, where the rubber meets the road in terms of transforming data into information, the integrity of which is mission critical.

In fact, businesses are finding that ETL has inherent inefficiencies as the amount of data being extracted from various data sources increases. The Extract step of ETL becomes resource intensive and costly, and impacts on the availability of data become more significant. For most businesses, the Extract, Transform and Load steps happen in batches during periods of limited utilization. Increasingly, though, given the international scope of many organizations, there are no periods of limited utilization and the entire process introduces unacceptable delays and costs.

Similarly, while the applications and primary data warehouses themselves need to be highly available for transactional purposes, decision support, archiving, backups, disaster recovery, upgrades, and migrations can badly degrade the availability of these databases for transaction processing.

As a result, many organizations are now looking to leave behind full data extracts from various data sources and are turning instead to a “Change Data Capture” (CDC) technique. As its name implies, only data that has changed in the data sources moves to the Transfer and Load components of the data aggregation process. This results in drastically lower overhead, is far faster, and brings with it an inherent process for auditing and compliance concerns. By examining only the data that changes, businesses can quickly identify possible problems or regulatory issues, in addition to streamlining aggregation.
ETL serves two other important functions as well: replication and auditing — meaning that the replication of data across multiple copies of a warehouse or data store is necessary to ensure limited or no downtime and high availability. Again, by shifting to a CDC paradigm, replication is much faster and can even happen in near real time.

Being able to follow the transformation steps applied to data is essential to ensuring its integrity and troubleshooting data integration issues. Well-designed ETL processes enable this level of auditing and transactional awareness, while CDC ensures that widespread problems can be identified early, before thousands or even millions of problematic data records are loaded into the warehouse.

Because Oracle databases are so commonly used as transactional data stores and data warehouses, specific tools exist to not only streamline these processes in Oracle environments, but also to address issues of availability and concurrent access to Oracle products. For example, these tools can simultaneously capture and load data from Oracle data stores into both operational and transactional data warehouses or replicate Oracle data stores on the fly. Thus, one copy of a data warehouse may be devoted strictly to high-performance capture of transactions like online orders and point-of-sale data. The same CDC-TL paradigm can then create a copy of the data warehouse in near real time to support DSS and critical reporting without any performance hits for either application.

The same approach can support backups, seamless migrations to new servers, failover, load balancing, or upgrades to new versions of Oracle, again with no performance degradation in mission-critical Oracle databases handling transaction processing.
BEGINNING THE PROCESS – KNOW AND LOVE YOUR DATA

None of this, though – not the information synthesis, the storage and auditing, the quality assurance or integrity verification, the mining or analytics – can happen without the right information architecture.

We wouldn’t allow a surgeon with only a cursory understanding of human anatomy to operate on us — nor should businesses press ahead with analytics, BI, and other informatics without a crystal clear understanding of the data with which they’re working.

Information architecture is painful and often tedious work, filled with diagrams, extensive documentation, countless meetings among stakeholders and experts, and the willingness to establish clear goals and objectives for the use of each source of data. No one person will have all of the answers or absolute expertise on every system to be integrated (other than perhaps the smallest of small businesses). As with most things in
21st century business, collaboration and communication, straightforward requirements, and a clear vision are core elements of any data-related strategy. To tame the data beast, all of the stakeholders need to understand the sources of data, the end game for the data, and the sorts of information they are trying to abstract from the data.

**SUMMARY**

Without data and the tools to harness it, there is little to differentiate 21st century business from that of the 20th century, except perhaps our shiny new devices. For now, business leaders often talk about data streams. In the coming years, these streams will become rivers. Taking the analogy a step further, as these rivers dump into lakes, then seas, and finally oceans, our ability to navigate will be a far more important differentiator from competitors than virtually any other factor.

In particular, when organizations invest in high-availability Oracle databases, optimizing for performance and transactional processing, the data they store must be simultaneously available for decision support. Downtime is not an option and botched migrations, upgrades, and backups aren’t just inconvenient; they can cost an organization millions, thus making systems that improve fault tolerance and concurrent access as critical as the databases themselves. Ensuring continuous high availability is possible with powerful data replication solutions and innovative approaches to data management.

---

**About Dell**

Dell Inc. (NASDAQ: DELL) listens to customers and delivers worldwide innovative technology, business solutions and services they trust and value. Dell offers a variety of IT management software products, including SharePlex® for Oracle.

SharePlex for Oracle is a mature, high-performance, high-availability technology that offers a low-cost alternative to other Oracle database replication tools. Unlike other solutions, SharePlex provides data compare and repair, in-flight data integrity, plus monitoring and alerting functionalities – all in a comprehensive packaged solution.

For more information, visit [www.dell.com](http://www.dell.com) and [quest.com/shareplex-for-oracle/](http://quest.com/shareplex-for-oracle/).