StatSlice

Bringing Big Data into the Enterprise

Overview

When evaluating Big Data applications in enterprise computing, one often-asked question is how does Big Data compare to the Enterprise Data Warehouse (EDW)? What does Big Data bring to the organization that the EDW cannot handle?

This article presents a technical and business discussion on the Big Data question. The technical discussion assumes familiarity with data architecture. The business discussion draws some conclusions about the actual application of Big Data in the enterprise.

Big Data vs. the Enterprise Data Warehouse

Big Data hardware is quite similar to the EDW's massively parallel processing (MPP) SQL-based database servers. EDW vendors include Teradata, Oracle Exadata, IBM Netezza and Microsoft PDW SQL Server. Both Big Data and EDW SQL database servers are composed of a large racks of Intel servers (each server called a node) and both distribute data across the nodes. Each node has local hard drive(s) for data storage and does not use a centralized storage system such as a Storage Area Network (SAN) in order to prevent I/O contention.

The first major technology difference is that Big Data's most common software platform, known as Hadoop, is free, open source and runs on commodity (non-proprietary) hardware. Most EDW vendors use propriety hardware with additional hardware accelerators to allow the servers to work better as a SQL style relational database. These hardware costs, when combined with the EDW vendor's proprietary software, typically reach cost levels that are exponentially higher per Terabyte than when using a Hadoop big data platform. Hadoop has more limitations than a SQL relational database but is far more scalable at a much lower price.

While Hadoop and EDW databases break apart large data sets into massively parallel systems, the actual implementation is substantially different. EDW databases parallelizes the data across smaller logical SQL database that exist on each node. Data is imported via a loading process that divides the data into the logical databases on a row by row basis, based on a data key column.

This extract, transformation and loading (ETL) process typically does additional data cleansing and data homogenization that matches the data with existing data in the EDW.

In contrast, Hadoop locates source files across a distributed file system (DFS). Conceptually, each file represents a segment or a partition of a table and files are simply duplicated across three nodes for redundancy. Hadoop data processing then uses direct access to the files. In other words, source files can be copied "as is" into the DFS which adds additional metadata to the file to aid in efficient data retrieval. This loading process is simpler and faster than a typical EDW ETL load.

The result: Hadoop can process large volumes of data, assuming the source data files are in a readable and ready state. Hadoop also needs the data to be self-contained, with the database analyst not looking to "join" the data to other data tables as is typically done in a relational database. You can still join data together, particularly using HBase extensions, but you would not use a normalized data model that is considered a best practice in SQL database modeling. An EDW database may have hundreds or even thousands of tables. Hadoop requires massive de-normalization and may have as few as one or two tables.

Hadoop does not excel at merging data across nodes at the detail data level, only at the reduce stage. The reduce stage is analogous to merging summarized data. Hadoop requires what is called co-location of data at the same node. In other words, if data in file A joins with data in file B, then file A and B must exist on the same node. EDW databases do not have this limitation. While they will perform better with a data model and indexing that co-locates data, in a normalized data model it is usually impossible to co-locate all tables. As discussed earlier, EDW vendors have considerable technologies to allow for flexible data models.

Also of interest is the fact that Big Data has a much shorter development time to load data—but a longer development time to query the data. EDWs have long lead times to add data sources to the database, but the time to write SQL queries is relatively short. Hadoop's map-reduce query language is Java (with other scripting option) with HBase and/or Pig SQL-like extensions that can aid in development. These still require higher skills and more time consuming effort than simply writing SQL. Business intelligence front ends are also limited with Big Data as compared with the EDW. This weakness will improve over time as Hadoop matures as a product.

In summary, Hadoop's limitations with data modeling prevent its application as a data warehousing database. Hadoop methodologies are in many ways brute force and simplicity.

But it's that simplicity that make it applicable for certain purposes. Many large data sets exist in the enterprise that actually fit within its limitations. When that environment exists, Big Data can bring analytics and business intelligence (BI) at warp speed.

Increasing Financial Gains from Low Value, High Volume Data

There are basically two types of data being created by enterprises and their customers:

- Low value, high volume data
- High value, low volume data

The typical EDW contains **high value data** that stores accounting, finance, sales, inventory, and other operational data. Each record in the database has real dollars associated with it. An enterprise needs to track dollars flowing through its operations and wants an end-to-end integrated view of that data. An EDW uses considerable resources to load that data from the various operational databases so all the data can be analyzed together. The EDW has hundreds, if not thousands, of tables that are in some way related. Although an EDW may have terabytes of high value data, that volume is still relatively low compared to the amount of that for low value data. This ratio typically has a proportionality relative to the size of the enterprise.

Low value data often does not have direct dollars associated with it. It is data typically not sourced from production databases. Application designers have already identified the high value data and integrated it into the application databases. Low value data is external from that process. It is usually sourced from files created from an application, instrumentation, monitoring, industrial controls or other system logs.

Given the source of low value data being log files, one of the great features of Hadoop is that these log files plays well with Hadoop's file centric methodology described earlier. Log files are also typically self-contained with all related attributed on a single row. In other words, the data is within the denormalized Hadoop "data model" limitation. Examples of low value, high volume data include:

- Website analytics
- Mobile app tracking
- Photo and video archives
- Social data and networking usage
- Collection of internet data via web crawlers
- Data collection from various systems including security cameras, factory instrumentations or robotics, medical records

The value proposition for Big Data is that it's a relatively cheap system to analyze large volumes of low value data that would otherwise be too expensive to implement using an EDW system. This allows enterprises to gain additional sources of intelligence that would otherwise have poor ROI or a high risk to unproven value.

Companies should view Big Data as business unit in the same way they typically view the EDW as a business unit. Most of the management processes will be similar. The main difference are the database server itself, employee skills sets and of course the data sources.

Big Data Application Examples

There are countless possibilities for Big Data applications. Tremendous volumes of low value data are currently being created by most companies. Anything that can be electronically monitored and its data acquired is a candidate. Anything in the enterprise that has potential business value should be evaluated and considered. The bottom line is that Big Data technologies lower the bar of what is economically feasible to acquire and analyze. Listed below are a few possibilities:

Computer Usage Analysis: Log files from applications servers from larger websites or mobile apps.

Usage Data from Social Networking: Enterprises can track their brand trending, allowing quick responses to both positive and derogatory damage to the brand.

Web Crawling the Internet: Identifying and storing relevant information of an enterprise from the Internet. There are many variations of this concept. One example is corporate financial data that allows investors to identify desirable statistical scenarios for investing.

Security Camera Analysis: Convert video feeds into traffic movements and store in Big Data so that traffic patterns can be analyzed. This allows on premise analysis of customer activity beyond "market basket analysis" of an EDW which only looks at actual purchases. One can correlate actual customer purchases with actual customer paths through the store to determine optimal store layouts. There is existing video content analysis software to convert video to traffic movement.

Factory Automation Data: Collect data from factory robotics, industrial controls, sensory instruments, RFID, and so on.

Medical Equipment Instrumentation: Analyzing 3D imaging systems.

Scientific Instrumentation: Collect data from data intense scientific experiments in astronomy, atmospheric science, genomics, biogeochemical, biological and other complex scientific research.

GPS Tracking and Transportation: Tracking rolling stock and conditions, military surveillance, and so on.

Internet/Network Analysis: Track data passing through switches and routers.

This is just a small sample of some of the potential uses for Big Data analytics. Much like companies evaluate data sources for their EDW they will need to review data sources for Big Data analytics. ROI calculations will be different but the payoffs can be substantial. Adding data sources to Hadoop costs much less than increasing data in your EDW. Consider the following from a ROI perspective:

- Costs for the server are a fraction of EDW server costs
- Disk space cost is minimal on Hadoop
- ETL development requires less time

Summary

With lower costs, you have lower risks of failure. The company needs to be willing to experiment and see what business value can be found in low value data sources. The EDW is here to stay, but in today's market of dramatically increasing high volume, low value data, the Big Data platform for some applications will yield dramatic results. One data point—in a recent survey by Dice, a leading tech job site, analytics and big data jobs are now in their top 10 list. A year ago they hardly got a mention.

Authored by Fred Zimmerman

Fred Zimmerman, a Data Architect for the consulting firm StatSlice Systems (www.statslice.com) is a veteran of data warehouse, business intelligence and database solutions with over 17 years of experience at Fortune 500 companies. Fred has proven experience integrating innovative ideas with industry best practices with the end result being streamlined, scalable and versatile data solutions.

Fred has designed business intelligence solutions for Verizon, Walmart, WellPoint, Coca-Cola Enterprises, Bank One (now Chase), Shell Oil, Microsoft Consulting Services and EDS (now Hewlett Packard).

His numerous specialties include experience with data warehousing, business intelligence and analytics, Master Data Management, all large-scale databases plus Hadoop and Hbase, and most Business Intelligence and ETL platforms (Microsoft, OBIEE, MicroStrategy, and Informatica).

For More Information

For more information about StatSlice Systems products and services, call (214) 206-9290 or email us at info@statslice.com. Please visit us at http://www.statslice.com.

This article will be published in the May/June 2013 edition of Enterprise Executive *magazine.*