Leveraging a Big Data Analytics Engine for Meaningful Insights

Keith B. Evans, Product Management, Gravic, Inc.
Paul J. Holenstein, Executive Vice President, Gravic, Inc.

The amount of information being generated each year is exploding at an unprecedented rate.[1] It is estimated that 80% of all of the world's data that has ever been created was produced in the last two years, and this rate is increasing. Social media such as Twitter and Facebook, articles and news stories posted online, blogs, emails, YouTube and other videos – they all contribute to what is now called big data.

Big data allows companies to obtain real-time business intelligence (RTBI) that they could never access in the past from their typical internal systems. Think of the customer-sentiment analysis that can be obtained simply from tweets. However, big data is a collection of data sets so large and so complex that it becomes impossible to process with current database-management tools and data-processing applications.

Much (perhaps most) of the content of big data is noise. It has little or no value to an organization. However, buried in this noise are tidbits of invaluable data which may be used to determine what customers are thinking, to plan new products, to find the strengths and weaknesses of competitors, to monitor for fraud and cyber-attacks, to defend against terrorism, and for many other purposes. The challenge is extracting the meaningful data from the noise. This is the task of the big data analytics engine.

A big data analytics engine typically requires a large network of tens, hundreds, or even thousands of heterogeneous, purpose-built servers, each performing its own portion of the task. All of these systems must communicate with each other in real time. They must be integrated with a high-speed, flexible, and reliable data-distribution and data-sharing backbone.

[1] This article is an update to our previous article, *Crunching Big Data*, which was published in the March/April 2014 issue of The Connection,

In this article, we look at several technologies that interact to extract valuable business information from the noise of big data.

Big Data

Events are no longer sufficient.

What do we mean by the above statement? After all, business processes and business intelligence are based on events. What did a customer purchase? When was a call made? When was an order delivered? Who logged on to our system and when?

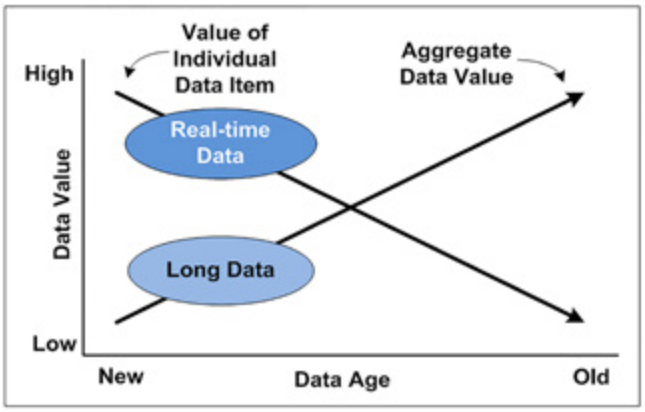

Figure 1 – The Value of Data Changes Over Time and Type

For decades, businesses have captured these events and stored them in transactional databases managed by highly reliable systems. Events drive the business. They determine production schedules, product deliveries, product re-order thresholds, banking, fraud detection, corporate financial statements, and a myriad of other business functions. A business would be paralyzed without its mission-critical online transaction-processing systems.

However, the world has evolved. The amount of available information that can be valuable to a company has rapidly expanded. Tweets, Facebook postings, news articles and newscasts, YouTube videos, the email and customer service calls a business receives – all of them may contain information advantageous to a company for making informed decisions and enhancing competitiveness. This is what we mean by big data – all of this data no matter the source or format.

The data stored in transactional databases represents high-density information. Every element is pre-determined to be important. However, transactional data is a tiny fraction of the total data generated worldwide. The data contained in big data is low-density. Most of the data is noise and has no real value to a company. But some of the data can be extremely important. How is the valuable data identified and extracted from the noise and put to use?

Real-Time Data and Long Data

There are, in fact, two types of big data that have to be managed – real-time data and long data.

Real-Time Data

Real-time data is used for immediate analytics and business decisions. The most immediate data available to a big data analytics engine is data that is streamed (pushed) to it, such as tweets, web clicks, emails, and customer calls. Other real-time data must be pulled from its sources, such as Facebook posts, news stories, and blogs.

Real-time data is characterized primarily by velocity and variety. Real-time data arrives in a variety of formats, and the big data analytics engine must be able to parse and process all of the real-time data that is presented to it with minimal processing delays.

Long Data

Long data is a massive data set that extends back in time over an extended period, such as over the life of an organization, and is important because it places real-time data in its proper perspective. If an organization does not look at events from an historical viewpoint, it will analyze current events as the norm and will be blinded to what has happened previously. It will miss repeated or unusual events and the opportunities (or threats) thereby presented. This perspective is why the addition of long data to an organization's source of information is so important. It provides context for current events.Consider climate change, for example.

Real-time data tells us that our ice caps are melting and that sea levels are rising. Is the culprit our increased carbon emissions, or is it a natural cycle that has gone on for eons? Long data can help answer this question.

The Time Value of Data

The value of information is a function of time. Interestingly, the relationship of value to time is opposite for real-time data and long data, as shown in Figure 1.

Real-time data is typically used for real-time decision making. The older it gets, the less useful it becomes. Some real-time data items may have half-lives of minutes. Others may have half-lives of microseconds (as is the case for algorithmic, high-frequency stock trading).

Long data, on the other hand, provides historical context for real-time data. The more historical data that is collected, the better is the context. Therefore, the value of long data increases over time as more and more data is accumulated.

The Big Data Analytics Engine

The general structure for a big data analytics engine is shown in Figure 2. As mentioned earlier, three types of data sources make up the information that flows into the analytics engine:

- Streamed data is pushed into the analytics engine, including sources such as Twitter and email. Stream processors are provided for each source to parse these streams and to deliver pertinent information to the analytics engine.

- Static data is pulled from other sources, such as Facebook postings and news stories. Fetchers are provided for each static source to fetch new data that has been added to that source and to deliver the fetched data to the analytics engine.

- Transactional data from the organization's transaction-processing systems is sent to the analytics engine for its real-time value as well as for its historical value.

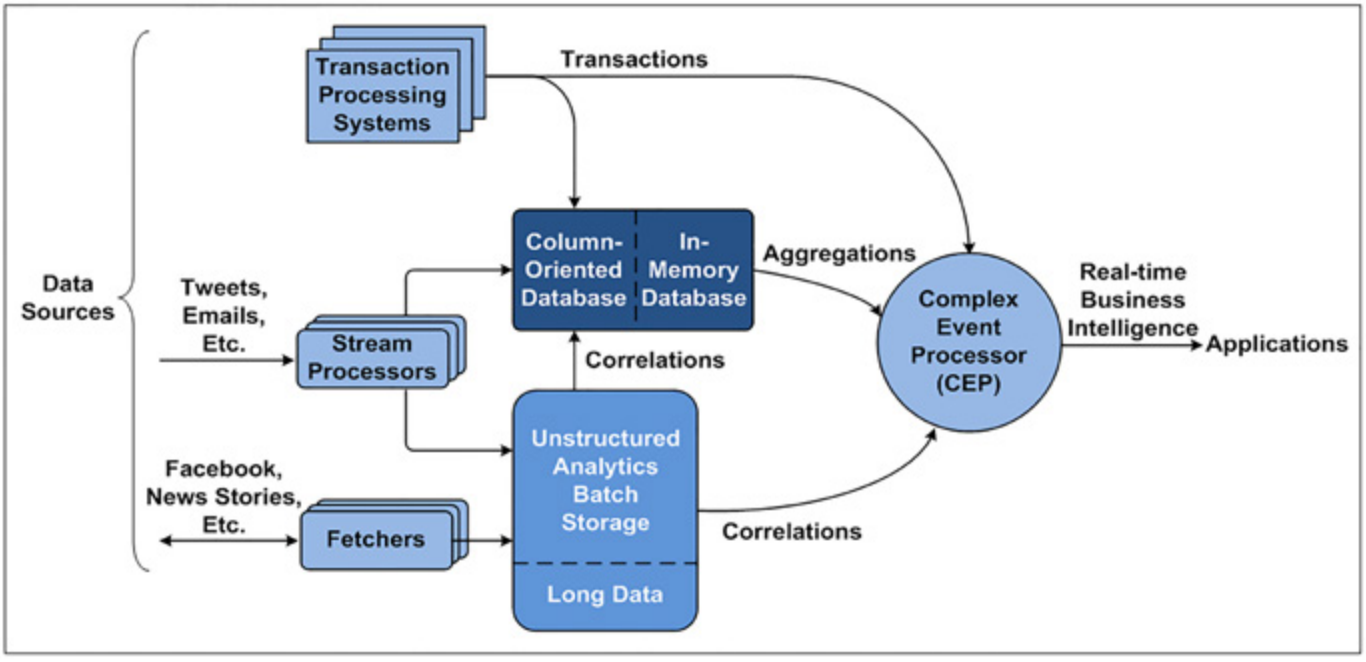

Figure 2 – Big Data Analytics Engine

• A batch-storage analytics engine is capable of storing unstructured data of any kind and can search that data rapidly for correlations. It receives all of the pertinent streaming data and all of the data that is being fetched from static sources. It analyzes this real-time data in context with the long data that it has stored to determine patterns of importance. These patterns, or correlations, are sent to other elements of the analytics engine for analysis processing.

• A column-oriented database stores intermediate results. A column-oriented database stores relational tables as columns rather than as rows. Many types of queries deal only with one or a few columns of a row and can be satisfied much more rapidly and efficiently with this architecture.

• An in-memory database typically holds the contents of the column-oriented database. It improves performance by eliminating disk-seek and transfer times and can further significantly speed up queries. Coupled with the column-oriented database, complex analytic queries can be completed in real time.

• A Complex Event Processor (CEP) combines data from multiple event streams in real-time to create more encompassing events. These latter events are the RTBI generated by the big data analytics engine. They provide in-depth insight into what is happening in the business. The goal of the CEP is to identify meaningful events such as business opportunities or threats and to allow immediate responses through the applications that the CEP feeds.

The Integration "Glue" for the Big Data Analytics Engine

As described previously, a big data analytics engine comprises many different systems with different missions. Each system is implemented on a "best-fit" platform with a "best-fit" database manager. There may be a myriad of heterogeneous platforms, applications, and databases that make up the analytics engine. A powerful, flexible, fast, and reliable data distribution fabric is required to interconnect these systems. The Shadowbase data replication engine (www.shadowbasesoftware.com) from Gravic, Inc. fulfills this role.

The data-distribution fabric between the many components in a big data analytics engine must be low-latency and provide high-capacity. It must be fundamentally heterogeneous and be able to deal with any application or database as a source or as a target. It must be able to reformat and restructure data on the fly as it moves data from one source to a totally different target. It must be highly reliable. All of these are attributes of Shadowbase software solutions.

The Shadowbase process-to-process architecture eliminates disk-queuing points that can slow down information delivery. Sub-second replication latency is achieved. The Shadowbase architecture can be multithreaded, including the communication channels, so that any desired data-transfer capacity can be attained.

Shadowbase replication supports heterogeneity. It can receive data as it is generated from any supported application or database and can deliver it to any supported application or database, including support for filtering, redefining, and enriching the information in-flight to satisfy any target environment

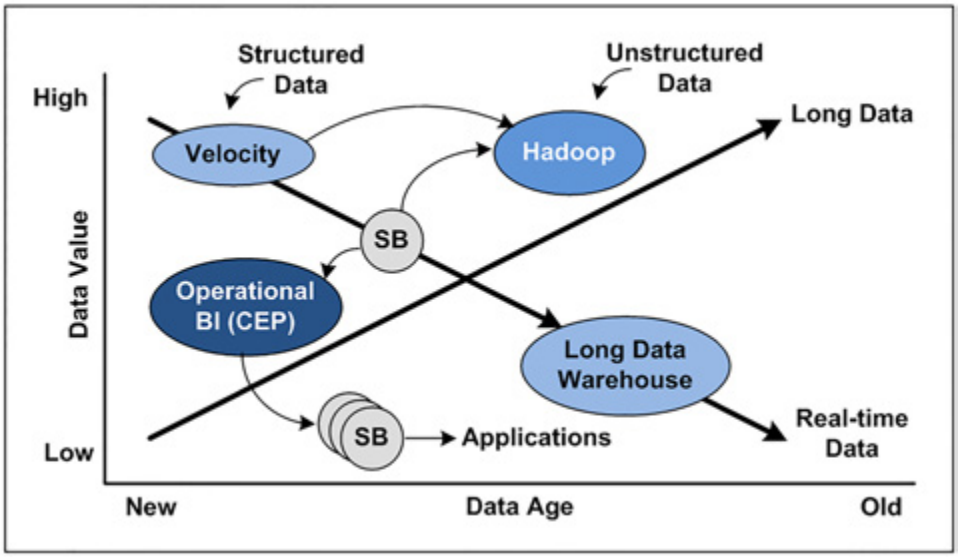

Figure 3 – Shadowbase (SB) as the Data-Distribution Fabric for Big Data Environments

formatting needs. Shadowbase software is architected to provide continuous availability. If one of its components fails, it is automatically restarted. Replication continues uninterrupted. If the target system fails, the Shadowbase engine queues all events until the target system is restored to service. It will then drain its queue of saved events to bring the target system back into synchronization with the data source and will automatically resume replication of real-time events.

An overview of using Shadowbase replication in a big data analytics engine is shown in Figure 3. This figure illustrates the age relationship between real-time data and long data, the functional components of the analytics engine, and how Shadowbase provides the data-distribution fabric between those components.

The stream processors and fetchers are represented by the "velocity" component. The CEP is the Operational Business Intelligence (OBI) component. Shadowbase technology can replicate structured data from the stream processor and fetcher databases to the CEP for BI analysis. Additionally, Shadowbase software can replicate the database change results of the operational BI analysis to downstream applications for additional processing.

A massive, unstructured database engine such as Hadoop (Google's database engine) provides storage for unstructured big data and analyzes that data. The stream processors and fetchers feed their data streams directly to Hadoop, and the Shadowbase replication engine feeds transaction data to Hadoop, which generates correlations and deduced events that are unstructured. By adding a custom capability to convert these unstructured outputs to a structured data stream, the Shadowbase engine can replicate the Hadoop correlations and deduced events to the CEP for real-time analytics. If a long data warehouse is available outside of Hadoop, the Shadowbase engine can replicate the incoming stream events, the fetched events, and the Hadoop-deduced structured events to the long data warehouse. Finally, Shadowbase technology can replicate the resulting business information to the enterprise's applications.

GRAVIC® Shadowbase

Summary

Big data offers the opportunity for businesses to obtain RTBI that they could never reach in the past from their typical internal systems. A big data analytics engine can mine social media, the press, email, blogs, videos, and a variety of other data sources to determine what customers are thinking, to plan new products, to find the strengths and weaknesses of competitors, to monitor fraud and cyber-attacks, to gain competitive advantage, and for many other purposes.

Big Data Protection

Businesses today are driven by data, and the quality of the business depends upon the quality of that data. Consequently, data has become one of a company's most valuable assets, and other people want it. Stealing or corruption of this data can result in significant business losses, pose serious security threats, and result in regulatory violations. As hackers become increasingly sophisticated, protection of data from unauthorized access is a number one priority for any IT department.

Protection of data from unauthorized access within a big data environment becomes much more complicated because the data is being consumed from many sources (trusted and otherwise), and moved between systems for analysis. For example, a data source may be restricted to only a certain set of users; if this data is replicated to a big data repository, measures must be taken to ensure that access to the data remains restricted to only this set of users.

Fortunately, using a data replication engine (such as Shadowbase) for the data distribution fabric addresses many of these data protection issues. When the data is in motion (being copied between systems) techniques such as IPSec and/or proxy servers (SSL/TLS) can be used to authenticate and encrypt each data packet. For data at rest, as the replication engine applies the data to the big data repository, Shadowbase user exits can be customized to encrypt or obfuscate the data as it is written. Encrypted target file systems can also be used when available. Via these means, the data replication engine ensures that data replicated to a big data repository remains protected, regardless of the source of the data.

Shadowbase replication capabilities can play a significant role in delivering inputs and outputs to key processes for analyzing big data. Wherever there is a need to transfer data from a data source to another target, regardless of the nature of those devices, Shadowbase software solutions can be placed into service to get the job done efficiently and reliably.

Keith B. Evans works on Shadowbase business development and product management for the Shadowbase product suite, including business continuity, data integration, application integration, zero downtime migration, data utilities, and synchronous replication, a significant and unique differentiating technology. To contact the author, please email: **SBProductManagement@gravic.com**.

Paul J. Holenstein is Executive Vice President of Gravic, Inc. He is responsible for the Shadowbase suite of products. The Shadowbase replication engine is a high-speed, unidirectional and bidirectional, homogeneous and heterogeneous data replication engine that moves data updates between enterprise systems in fractions of a second. It also provides capabilities to integrate disparate operational application information into real-time business intelligence systems. Shadowbase Total Replication Solutions® provides products to leverage this technology with proven implementations. For further information regarding Shadowbase data integration and application integration capabilities that can assist in solving big data integration problems, please refer to the companion documents **Shadowbase Streams for Data Integration** and **Shadowbase Streams for Application Integration**, or visit **www.ShadowbaseSoftware.com** for more information. To contact the author, please email: **SBProductManagement@gravic.com.**