

Every minute, there are 98,000 tweets, 695,000 Facebook postings, 700,000 Google searches, and 166 million emails, and the rate is increasing – and quickly. All of these data streams may contain information that can be extracted in real-time and can be correlated with information from other sources to provide immediate and valuable insights for an enterprise.

A large brokerage firm may want to scrutinize in real time five million trades per day as it looks for excessively risky trading by rogue brokers. A large mobile phone operator may want to analyze in real time 500 million call records per day to predict channel utilization. Real-time information delivery is one of the defining characteristics of big data analytics. Latency in the information pipeline, from data acquisition to data transfer, data storage, data analysis, and disseminating analytic results, must be kept to a minimum.

- **Variety** is the range of data types and sources that an analytics system will need to understand.

Data may arrive in different ways. Some data may be streamed, pushing its way into the big data analytics engine. Tweets and emails are examples of streaming data sources. Other data, such as Facebook postings and news stories, may have to be pulled from their sources by the big data analytics engine.

Until the advent of big data, corporate data was typically structured and stored in relational databases such as SQL databases. The databases are organized according to well-defined metadata. This paradigm changes with big data. Most big data information is unstructured or loosely structured. Information sources vary widely in the presentation of their data. Sources of information that an organization might utilize above and beyond the structured data of today include web-click streams, social media (Facebook, Twitter, blogs), customer behavior (PayPal, Google Wallet), geo-location (mobile phone GPS), audio (broadcast news), video (YouTube and live video feeds), digital pictures, and sensor readings for a variety of purposes. Though some of these data sources may have some useful structure, e.g., tweets and web clicks, others such as blogs and broadcast news have no structure.

Some practitioners have added additional Vs:

- Value is the worth of the data to an organization for such real-time functions as fraud detection, risk management, and targeted marketing. Many times, there is a temporal attribute added. Operational data (considered to be recent) is useful for tactical/analytic activity, whereas archival data (considered more historical) is useful for strategic/analytic activity. Which is more valuable? It depends on your goal.
- Veracity is the trust that business leaders place in the results they obtain from big data analytics..

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, transactional data from an operational system (for example, when sent by a data replication engine), 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.

Long data is typically built from real-time data and is important because it places the 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 2](#).

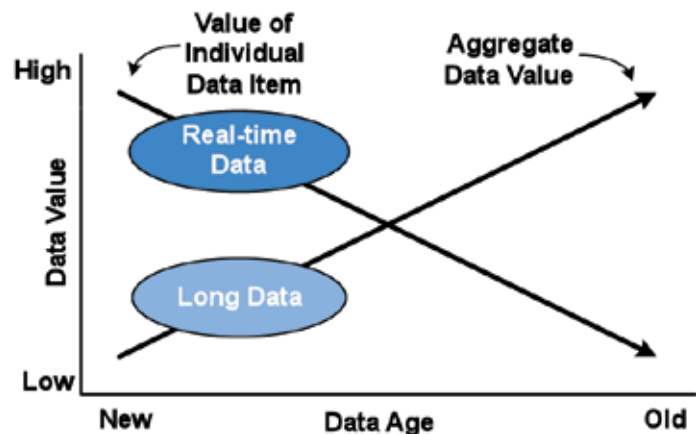


Figure 2 – The Value of Data Changes Over Time and Type

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 milliseconds (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 3](#). 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.

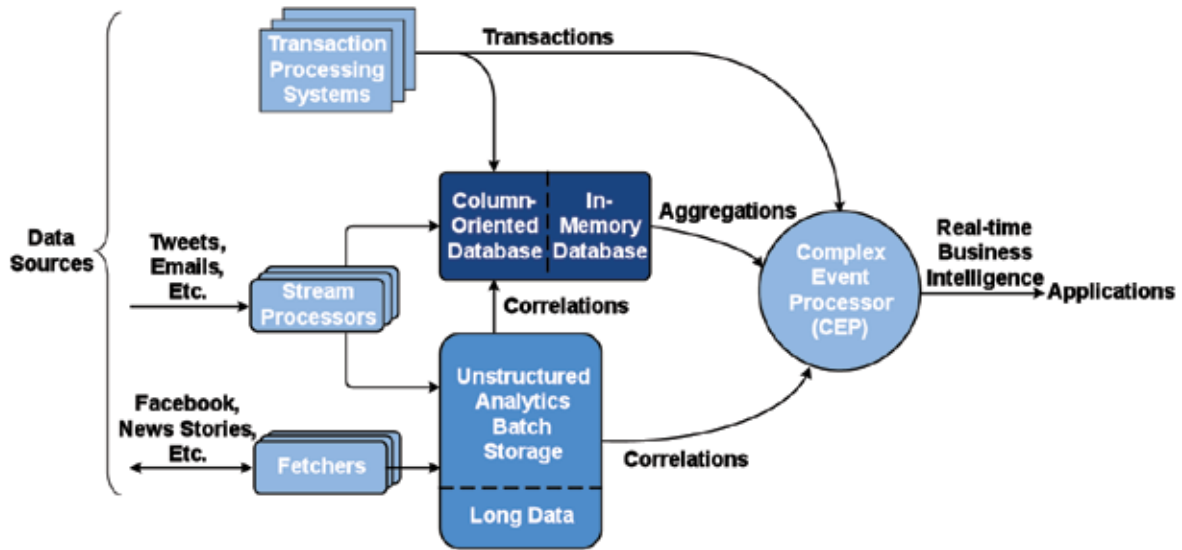


Figure 3 - Big Data Analytics Engine

- 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.

At the heart of the analytics engine are several (typically massive) components. The implementation of each can require tens or even hundreds of commodity servers:

- 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 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. In a row-oriented database (such as a classic relational SQL database), the entire set of rows must be read to obtain the information contained in just a few columns. Alternatively, indices can be set up to point to the data in columns; but creating and managing these indices is a CPU- and disk-intensive task. It typically is done for only a few columns deemed important to known applications with predetermined access needs. With all data stored by column rather than by row, queries against specific columnar data can be decided upon and executed much faster. For instance, the aggregation of values or the selection of a range of values can be accomplished just by reading the desired column data instead of potentially all of the rows in the table (which must then be culled down to the desired columnar information). When using a column-oriented database, claims have been made of query performance improving by a factor of two to three orders of magnitude compared to a classic row-oriented relational database.
- 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.gravic.com/shadowbase) 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 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

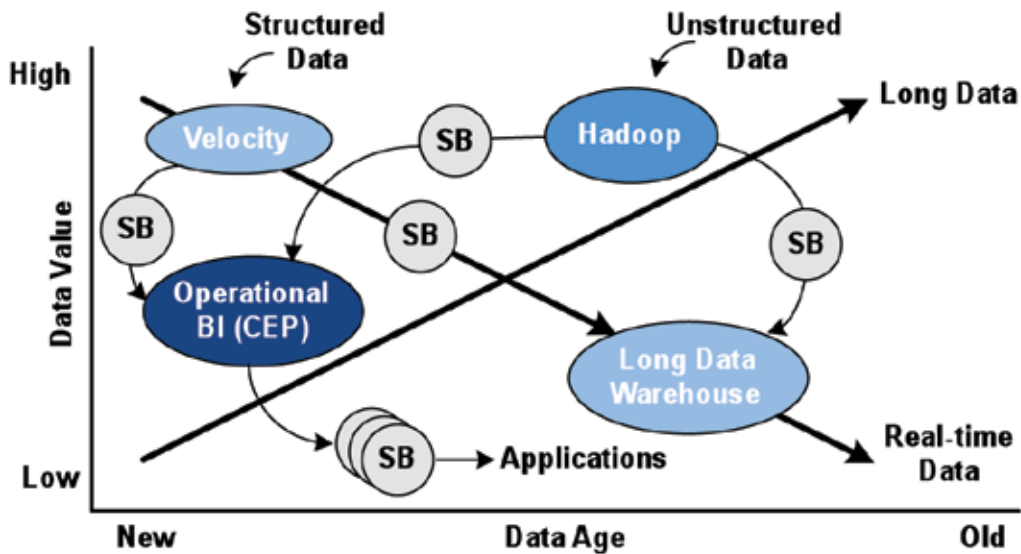


Figure 4 – Shadowbase (SB) as the Data-Distribution Fabric for Big Data Environments

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 4. 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. In this architecture, the CEP is represented by the Operational Business Intelligence (OBI) component. Shadowbase technology can replicate streams and fetched data from the stream processors and fetchers to the CEP for BI analysis.

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.

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, and for many other purposes.

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 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. 

References

- The following references are useful in exploring big data further: [Shadowbase Streams for Data Integration](#), Gravic white paper.
- [Shadowbase Streams for Application Integration](#), Gravic white paper.
- [Stop Hying Big Data and Start Paying Attention to ‘Long Data’](#), *Wired*; January 29, 2013.
- [HP, Dell Announce New Big Data Analytics Solution](#), *Data Center Knowledge*; February 28, 2013.
- [Big Data, Big Decisions](#), *Information Week*; April 1, 2013.
- [Big Data at the Speed of Business](#), *IBM White Paper*.
- [How Does Big Data Help You Compete and Innovate](#), *Mission Critical Computing Blog*; February 20, 2013.
- [Recent Cases of Big Data Where They Were Least Expected!](#), *Mission Critical Computing Blog*; February 21, 2013.
- [NonStop and the road to Big Data](#), *Mission Critical Computing Blog*; February 28, 2013.
- [Big Data](#), *Wikipedia*.
- [Column-oriented DBMS](#), *Wikipedia*.
- [Complex Event Processing](#), *Wikipedia*.
- [Content Analysis](#), *Wikipedia*.
- [Autonomy Corporation](#), *Wikipedia*.

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.Gravic.com/Shadowbase for more information. To contact the author, please email: SBProductManagement@gravic.com.