HPE Shadowbase Solutions in a Big Data World

A Gravic, Inc. White Paper

Executive Summary

The amount of information being generated each year is exploding at an unprecedented rate. It is estimated that 90% of the world's data was generated 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 are all contributing to what is now called *big data*.

Big data offers the opportunity for companies to obtain real-time business intelligence that they could never reach 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, comprising both structured and unstructured data, that it becomes impossible to process with current database management tools and data processing applications. Sometimes the term *open data* is used to describe big data that is publicly available for use and distribution.

Most big data's content is noise. It has 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 competitor strengths and weaknesses, to monitor for fraud and cyber-attacks, to defend against terrorism, and for many other purposes. The challenge is extracting meaningful data from the noise. This is the task of the big data analytics engine.

A big data analytics engine requires a large network of tens, hundreds, or even thousands of heterogeneous, purpose-built servers, each performing its own subtask. All of these systems must intercommunicate with each other in real-time. They must be integrated with a high-speed, flexible, and reliable data distribution and sharing backbone. The HPE Shadowbase software suite provides the ideal solution to this problem.

The Shadowbase data replication engine is the glue that binds the servers comprising a big data analytics system. Shadowbase replication reliably reformats and transfers large amounts of data between heterogeneous databases and applications in real-time, and helps users capitalize on the big data explosion. In this paper, discover how several technologies interact to extract and differentiate valuable business information from the noise, and the key role HPE Shadowbase software plays in these solutions.

> ***"Big Data: The next frontier for innovation, competition, and productivity."***
> — McKinsey Global Institute

Table of Contents

Executive Summary **2**

Big Data **4**
Events are no Longer Sufficient *4*
The Three Vs of Big Data *4*

Real-Time Data and Long Data **5**
The Time Value of Data *6*

The Big Data Analytics Engine **6**

HPE Shadowbase Data Replication is Integration "Glue" for the Big Data Analytics Engine **8**

Big Data Protection **8**

Case Studies **9**
Stock Exchange and Clearinghouse Gain Operational and Availability Benefits with HPE Shadowbase Solutions *9*
Large Telco Deploys HPE Shadowbase for Subscriber Data Mining and Real-time Fraud Detection *9*
Large International Tour Operator Utilizes Commodity Servers to Optimize Look-to-Book Processing and Implement ODS *9*
Real-Time Credit and Debit Card Fraud Detection: An HPE Shadowbase Real-Time Business Intelligence Solution *9*

Summary **10**

International Partner Information **11**

Gravic, Inc. Contact Information **11**

Table of Figures

Figure 1 – Measuring Data by the Byte 4
Figure 2 – The Value of Data Changes Over Time and Type 6
Figure 3 – Big Data Analytics Engine 7
Figure 4 – HPE Shadowbase (SB) Data Distribution Fabric for Big Data Environments 10

HPE Shadowbase Solutions in a Big Data World

Big Data

Events are no Longer Sufficient

What do we mean by this statement? After all, events drive business processes. They determine production schedules, product deliveries, banking, fraud detection, corporate financial statements, and a myriad of other business functions. Business processes and business intelligence are based on real-world events and help paint the picture of who did what, when, and where:

- *Who* made the purchase?
- *What* did they buy?
- *When* did they buy it?
- *Where* was it delivered?

For decades, businesses have captured these events and stored them in transactional databases managed by highly reliable systems.[1] A business would be paralyzed without its mission-critical online transaction processing systems.

However, as per usual, the world and its technology kept evolving. The amount of information being generated that can be valuable to a company has rapidly expanded beyond simple events. Social media, news articles, podcasts, videos, emails, phone calls – all of these mediums[2] may contain information invaluable to a company for making informed decisions and enhancing competitiveness. All of this information and the real-world events mentioned before is big data, no matter the source or format, structured or unstructured.

Data stored in transactional databases represent high-density information. It is a tiny fraction of the total data generated worldwide, but every element is important. However, data contained in big data is low-density. Most of the data is noise and has no real value, but some may be extremely important. How can a company identify and filter the noise, leaving valuable data that can be put to good use? Let's explore the solution to this challenge.

The Three Vs of Big Data

Big data is commonly characterized by the three Vs – *volume*, *velocity*, and *variety*:

1. **Volume** is *the amount of data that is generated that may need to be analyzed by an analytics system.*

   Data volume available to businesses today is staggering. For instance, every day:
   - over thirty exabytes (10^{18}) are generated worldwide,
   - twelve terabytes of tweets can be mined for customer-sentiment analysis (what do people really think of you?),
   - and nearly 1 million electric meter readings can be used to better predict power consumption and set variable rates to soften peak loads.

   Even so, petabyte and exabyte databases are about to become old news, with Brontobyte databases just around the corner! (See **Error! Reference source not found.** to get an idea on size.)

2. **Velocity** is *the rate at which data is entering the analytics system.*

Kilobyte	10^3
Megabyte	10^6
Gigabyte	10^9
Terabyte	10^{12}
Petabyte	10^{15}
Exabyte	10^{18}
Zettabyte	10^{21}
Yottabyte	10^{24}
Brontobyte	10^{27}

Figure 1 – Measuring Data by the Byte

[1]These databases typically contain *structured data* such as a field that conforms to constraints based on its datatype.
[2]These mediums typically contain *unstructured data,* meaning they do not conform to predefined constraints and/or organization.

Data velocity is also increasing at a staggering rate. For instance, every day:

- 500,000,000 tweets,
- 695,000 Facebook posts,
- 700,000 Google searches,
- and 166 million emails are sent.

All of these data streams may contain information that can be extracted in real-time and correlated with information from other sources to provide immediate and valuable insights for an enterprise.

A large brokerage firm may want to scrutinize five million trades per day in real-time to identify excessively risky trading by rogue brokers. A large mobile phone operator may want real-time analysis of 500 million call records per day to predict channel utilization.

Real-time information is one of the defining characteristics of big data analytics. Latency in the information pipeline, from data acquisition to transfer, storage, analysis, and disseminating analytic results must be kept to a minimum.

3. **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 streaming, pushing its way into the big data analytics engine. Tweets and emails are examples of 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 database formats such as SQL. Relational 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, other sources such as blogs and broadcast news have no structure.

4. **Other Vs**

   Some practitioners have added additional Vs such as *value* and *veracity*:

   A. **Value** is the *data's usefulness to an organization*.

      Some real-time examples include fraud detection, risk management, and targeted marketing. Typically, there is a temporal attribute added, as in operational data (considered more recent), which 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.

   B. **Veracity** is *the trust that business leaders place in the results they obtain from big data analytics.*

      This trust is partly based on the veracity of the original data. Remember the adage from the early days of computing, "garbage in, garbage out." This means the value of information derived from analytics is dependent upon the original value of the input data.

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*.

A. **Real-time data** is *data used for immediate analysis to quickly make business decisions*.

Streamed data provides the most immediate and relevant data to a big data analytics engine, such as tweets, web clicks, emails, and customer calls. Other real-time data must be pulled from their 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.

B. **Long data** is a *massive dataset over a long period of time, such as over the life of an organization.*

Long data is important because it helps paint a picture from the data's lifetime, adding perspective to real-time data. If an organization does not look at events from a historical viewpoint, it will analyze current events as the norm and be blinded to what has previously occurred, missing repeated or unusual events and the opportunities (or threats) 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. For example, consider the debate about the cause of climate change. Real-time data tells us that the world's ice caps are melting and sea levels are rising. Is the culprit increased human carbon emissions, or is a natural cycle that has occurred for eons? Long data can be used to help answer this question.

The Time Value of Data

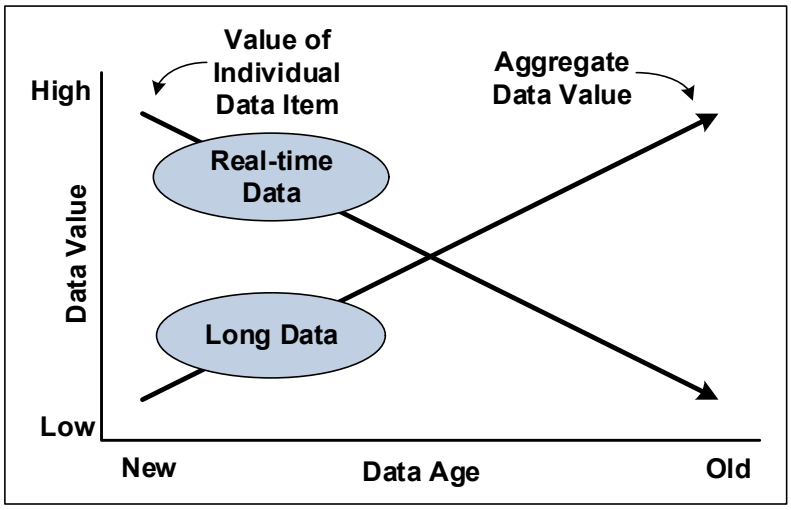

Figure 2 – The Value of Data Changes Over Time and Type

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.

Real-time data is typically used for real-time decision making. As data ages, its usefulness fades away. Some real-time data may have half-lives of minutes. Other data 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 collected, the better the context. Therefore, long data increases in value as it accumulates.

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 comprise the information that flows into the analytics engine:

- Streaming 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.

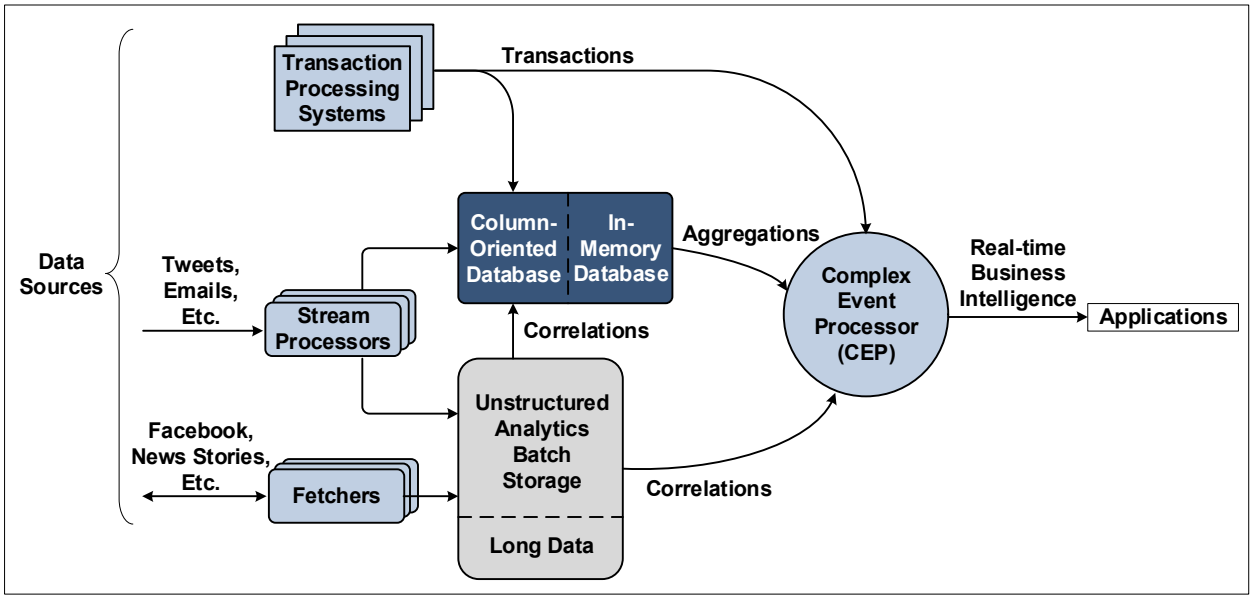

Figure 3 – Big Data Analytics Engine

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 analysis processing.

- A column-oriented database stores intermediate results. It also stores relational tables as columns rather than 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. This 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 real-time business intelligence 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.

HPE Shadowbase Data Replication is Integration "Glue" for the Big Data Analytics Engine

- Shadowbase connects heterogeneous systems, databases, and applications within big data analytics engines, eliminating data silos and ensuring different teams and departments operate with unified, consistent data.
- It delivers high-speed, low-latency data replication through a multithreaded, process-to-process architecture that bypasses disk queues and supports sub-second latency, enabling fast, reliable communication between systems.
- Shadowbase can replicate data from any supported source to any supported target, whether it's an application, a database, or a log file, and can apply filtering, redefinition, and enrichment to the data while it's in motion to match the target format.
- It handles complex data transformations across diverse environments, offering built-in reformatting capabilities and user exits for defining custom algorithms when unique restructuring is needed.
- Supports both asynchronous and synchronous replication, allowing flexibility depending on application needs. Both methods ensure transactional consistency and keep the target system available for processing.
- Designed for continuous availability, Shadowbase automatically restarts failed components, queues data during target outages, and resynchronizes once the target is restored—without interrupting real-time replication.
- The software is optimized for big data architectures, offering low latency, high throughput, powerful message processing, flexible integration endpoints, and strong support for heterogeneous environments and continuous uptime.

Integrating heterogeneous data from different sources is a formidable challenge, one that has been solved by Shadowbase data replication, which provides the ideal foundation as the data distribution fabric for a big data analytics engine.

Big Data Protection

Businesses today are driven by data. The quality of services provided depends upon the quality of business data. Consequently, data has become one of a company's most valuable assets, now hailed as "the new currency," and malicious actors want it. Data theft or corruption can result in significant business losses, pose serious security threats, and result in regulatory violations. As hackers and malicious actors become increasingly sophisticated, protection of data from unauthorized access has become a top priority for IT departments.

HPE Shadowbase Data Protection

Case Studies

In the following section, we provide examples of the ways in which customers are taking advantage of Shadowbase capabilities for data integration to benefit their business and customers.

Stock Exchange and Clearinghouse Gain Operational and Availability Benefits with HPE Shadowbase Solutions

***The Connection,* by Paul J. Holenstein**
A South American stock exchange using a trading system built with HPE NonStop Servers faced high data..

Large Telco Deploys HPE Shadowbase for Subscriber Data Mining and Real-time Fraud Detection

A large telephone company (telco) switched from providing landline services to mobile phones and began using HPE Shadowbase data replication solutions to keep billing services up-to-date and to prevent cellular fraud...

Large International Tour Operator Utilizes Commodity Servers to Optimize Look-to-Book Processing and Implement ODS

This case study is an excellent example of HPE's move to a converged infrastructure. It demonstrates how one customer took advantage of facilities available to the HPE community to integrate diverse systems into a common application...

Real-Time Credit and Debit Card Fraud Detection: An HPE Shadowbase Real-Time Business Intelligence Solution

A major provider offers inter-bank transaction-switching services by capturing information during a point-of-sale – at a store terminal or website – or at an ATM. A complex suspicious or fraudulent activity determination is made...

Summary

Big data offers an opportunity for businesses to obtain a level of real-time business intelligence (RTBI) that they could never reach in the past from their typical internal systems. A big data analytics engine can mine social media, news articles, 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.

A big data analytics engine mines this data regardless of its the volume, velocity, and variety. It is a highly complex mixture of different, built-for-purpose systems that must be integrated with a high-speed, flexible, and reliable data distribution backbone such as HPE Shadowbase solutions (see Figure 4).

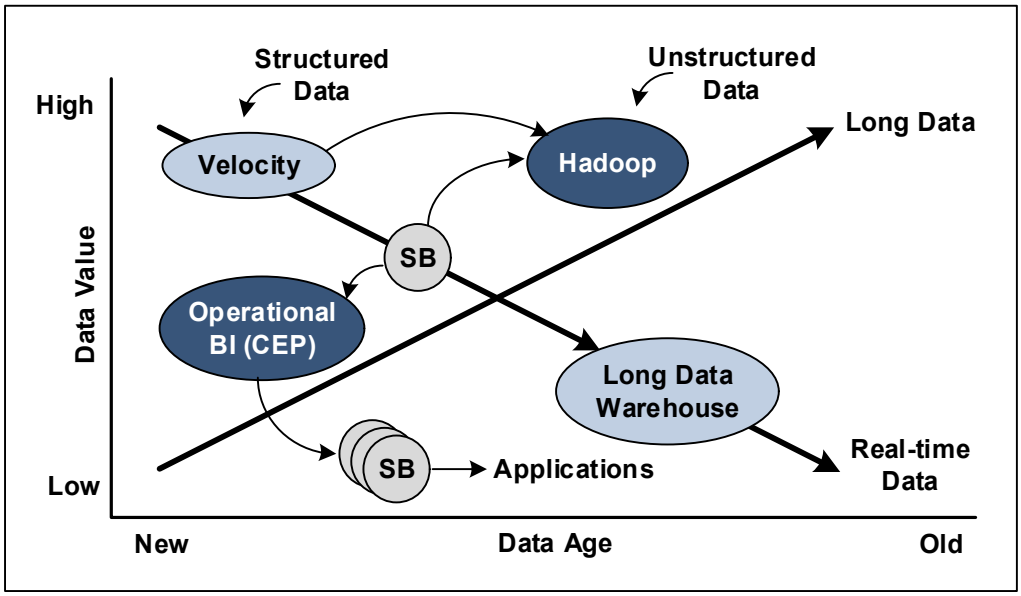

Figure 4 – HPE Shadowbase (SB) Data Distribution Fabric for Big Data Environments

Figure 2 indicates the age relationship between real-time and long data, and Figure 4 expands on this by including functional components of the analytics engine. The stream processors and fetchers are represented by the "velocity" component. The CEP provides 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 and to trigger business actions.

Figure 4 is only one example of the powerful capabilities Shadowbase replication[3] can play in a big data environment. 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.

[3]For more information, see *HPE Shadowbase Streams for Data and Application Integration*.

International Partner Information

Global

Hewlett Packard Enterprise
6280 America Center Drive
San Jose, CA 95002
USA
Tel: +1.800.607.3567
www.hpe.com

Japan

High Availability Systems Co. Ltd
MS Shibaura Bldg.
4-13-23 Shibaura
Minato-ku, Tokyo 108-0023
Japan
Tel: +81 3 5730 8870
Fax: +81 3 5730 8629
www.ha-sys.co.jp

Gravic, Inc. Contact Information

17 General Warren Blvd.
Malvern, PA 19355-1245
USA
Tel: +1.610.647.6250
Fax: +1.610.647.7958
www.shadowbasesoftware.com
Email Sales: shadowbase@gravic.com
Email Support: sbsupport@gravic.com

Hewlett Packard Enterprise Business Partner Information

Hewlett Packard Enterprise directly sells and supports Shadowbase Solutions under the name ***HPE Shadowbase***. For more information, please contact your local HPE account team or visit our website.

Copyright and Trademark Information

This document is Copyright © 2014, 2017, 2020, 2023, 2025 by Gravic, Inc. Gravic, Shadowbase and Total Replication Solutions are registered trademarks of Gravic, Inc. All other brand and product names are the trademarks or registered trademarks of their respective owners. Specifications subject to change without notice.