

# **Table of Contents**

- **Executive Summary**
- **The 3 V's (Volume, Velocity, Variety) of Big Data**
- **Harnessing "Big" Data Sets – Big Data Analytics**
- **Key Ingredients of Big Data Analytics**
- **Use Cases for Big Data Analytics**
- **Conclusion**
- **Appendix**

## Executive Summary

Every day, every second, huge volumes of data are being created as a by-product of business and individual activity. Companies are tapping seemingly endless streams of data about their customers, suppliers, and operations, while sensors — which increasingly are embedded in devices such as mobile phones, smart energy meters, automobiles, and industrial machines — are also recording and communicating data. Contributing further to this deluge of data is the use of social media sites, smart phones, and other consumer devices (including PCs and laptops), by billions of people around the world.

The explosive growth in the amount of data created in the world continues to accelerate and surprise us in terms of sheer volume. The data deluge is happening everywhere and is not only restricted to niche sources. It encompasses sensor and machine data, transactional data, metadata, and social network data as evidenced by the examples highlighted below.

*Boeing jet engines can produce 10 terabytes of operational information for every 30 minutes they turn. A four engine jumbo jet can create 640 terabytes of data on just one Atlantic crossing; multiply that by the more than 25,000 flights flown each day, and you get an understanding of the enormous amount of data produced<sup>1</sup>.*

*Transactional data has grown in velocity and volume at many companies. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases more than 2.5 petabytes of data—the equivalent of 167 times the books in America's Library of Congress<sup>1</sup>.*

*Social network data is another source adding to the superabundance of data. The micro-blogging site Twitter serves more than 200 million users who produce more than 90 million "tweets" per day, or 800 per second. Each of these posts is approximately 200 bytes in size. On an average day, this traffic equals more than 12 gigabytes and, throughout the Twitter ecosystem, the company produces a total of eight terabytes of data per day<sup>1</sup>.*

*Facebook announced they had surpassed the 750 million active-user mark, making the social networking site the largest consumer-driven data source in the world. Facebook users spend more than 700 billion minutes per month on the service, and the average user creates 90 pieces of content every 30 days. Each month, the community creates more than 30 billion pieces of content ranging from Web links, news, stories, blog posts and notes to videos and photos<sup>1</sup>.*

Everywhere you look, the quantity of information in the world is soaring. The term "**big data**" has emerged to describe this monstrous growth in data. Big Data represents data sets whose characteristics are comprised of high volume, high velocity and a variety of data structures.

## **The 3 V's (Volume, Velocity, Variety) of Big Data**

**Volume**: Data Volume is the primary attribute of Big Data. Volume is often quantified in terms of terabytes of data. Anything between 3 – 10 terabytes of data falls within the realm of “Big Data”. In addition, data volume can also be quantified by counting records, transactions, tables and files. A large number of records, transactions, tables or files can be categorized as “Big Data”. Volume of data is one of the defining characteristics of “Big Data”; however, data velocity and data variety (highlighted below) constitute the other key characteristics/ingredients of “Big Data”.

**Velocity**: Speed or Velocity of data is another defining characteristic of “Big Data”. Data Velocity encompasses the frequency of data generation and the frequency of data delivery. In today’s hyper connected and networked society, there is a continuous stream of information coming from a range of devices ranging from sensors, robotics manufacturing machines, video cameras to mobile gadgets. This ever increasing amount of data relentlessly flying from devices in real-time is causing data volumes to grow and get big in a hurry.

**Variety**: One thing that makes big data really big is that it’s coming from a greater variety of sources than ever before. Data from web sources (i.e. Web logs, clickstreams) and social media is remarkably diverse. RFID data from supply chain applications, text data from call center applications, semi-structured data from various business-to-business processes, and geospatial data in logistics constitutes an eclectic mix of data types that makes variety/diversity an important attribute characterizing “Big Data”.

## **Harnessing “Big” data sets – Big Data Analytics**

The above-mentioned characteristics of “Big Data” bring new challenges to data integration, information discovery and exploration, and reporting. Big data sets can no longer be easily managed or analyzed with traditional or common data management tools, methods and infrastructures. New techniques are required to harness this “Big Data” to dramatically improve decision making.

Fortunately, the exponential growth in the price-performance of computing, digital storage, and network bandwidth have converged to bring down the cost of collecting and analyzing these huge data-sets. Therefore, much of the large-scale data that’s long been collectable can now be stored, aggregated, and filtered cost-effectively.

Today, enterprises are exploring big data to discover facts they didn’t know before. This is an important task right now because the recent economic recession forced deep changes into

most businesses, especially those that depend on mass consumers. Using big data analytics, businesses can study big data to understand the current state of the business and track still-evolving aspects such as customer behavior.

At the center of the big data movement is an open source software framework called Hadoop which utilizes large number of computing nodes to capture valuable insights from big data sets and facilitate big data analytics. Hadoop utilizes MapReduce which is a powerful framework for processing data sets across clusters of Hadoop nodes. The Map and Reduce process splits the work by first mapping the input across the control nodes of the cluster, then splitting the workload into even smaller data sets and distributing it further throughout the computing cluster. This allows it to leverage massively parallel processing (MPP), a computing advantage that technology has introduced to modern system architectures. With MPP, Hadoop can run on inexpensive commodity servers, dramatically reducing the upfront capital costs traditionally required to build out a massive system. As the nodes "return" their answers, the Reduce function collects and combines the information to deliver a final result.

### **Key Ingredients of Big Data Analytics:**

The cornerstones to enabling Big Data Analytics is the ability to access and integrate information of any scale, from any source; combining transaction/structured data with unstructured data gleaned from social media sites/web properties/sensor information to enable insights not possible before; utilizing frameworks constructed for data-intensive processing running on a cluster of commodity hardware and ultimately feeding the processed results to a Business Intelligence (BI) environment that allows the ability to search, discover and visualize information through different perspectives so that an exhaustive analysis of the information is possible. Essentially, Big Data Analytics is based on 4 key pillars:

- Acquire and integrate data (regardless of scale) from any source
- Harness interaction data (i.e. social network data, device/sensor based RFID information) to enhance/enrich transaction data to enable unprecedented insights
- Utilize Massive Parallel Processing frameworks on commodity hardware to process big data sets encompassing transaction data, sensor/machine data and social network data
- Collect and Deliver the processed results to a BI environment for exploration purposes to derive precious, golden insights

## **Use Cases for Big Data Analytics**

Enterprises are collecting more data than ever before. Big data analytics provide new ways for businesses and government to analyze big data sets and discover new business facts that no one in the enterprise knew before. The value of Big Data Analytics is showcased by the use cases highlighted below:

**Customer Satisfaction Analytics:** Currently, there is no clear way of deciphering issues with products and the related customer satisfaction levels associated with those products. The inability to effectively marry structured with unstructured content and classify/group related product issues into “problem clusters” prevent enterprises from pro-actively monitoring customer satisfaction levels. Utilizing advanced analytical techniques on big data sets for customer analysis will enable enterprises to gain much-needed insight into customer’s mindset and help companies pro-actively detect problems and prevent the issues from recurring in current and future product lines.

**Competitive Analytics:** In today’s hyper-competitive environment, it is imperative for enterprises to have a constant pulse on the competition so that they can sustain/maintain their competitive advantage. By capturing vast amounts of competitive information from different web sites, social media sites and public domains and utilizing big data analytics on the information garnered from the different sources, enterprises can obtain critical insights into the strengths and weaknesses of their competitors and identify growth opportunities for their respective product lines.

**Healthcare Analytics:** Epidemics or seasonal health issues (i.e. influenza) is marked by certain symptoms which spread to a larger section of the population if not detected early and treated pro-actively. The symptoms usually vary among individuals and the treatment for the symptoms varies across doctors. Current problems with the common classification of symptoms and the appropriate treatments for the symptoms prevent doctors from effectively diagnosing and controlling the outbreak of epidemics. Adopting Big Data analytics to decipher the symptoms and identifying appropriate treatments for the symptoms will enable doctors to effectively tackle and control health issues before it impacts a larger section of the population.

**Product Usage Analytics:** Currently, product-focused companies lack the insight to understand *what* features of their products are being utilized and *how* the features of the product are being utilized (i.e. usage patterns). Utilizing Big Data analytics to detect product usage from machine-generated information and sensor-based RFID information can enable product-focused companies to effectively monitor, manage and direct product investments thereby generating a better return on Research & Development spend.

**Inventory Analytics:** Suppliers are continuously trying to determine methods to better or more effectively manage their supplies/stocks. Most suppliers today lack insight into the exact number of their products on every shelf of every store at precise moments in time. This

inability to garner a complete overview of *where* and *when* their products are selling leads suppliers to over stock supplies so that they can meet demand in a timely fashion. However over-stocking supplies leads to higher inventory costs which in turn impacts the suppliers bottom line. Adopting Big Data Analytics on sensor-based RFID information can equip suppliers with information about the rate (i.e. by hour, by day) at which their supplies are being sold across different stores so that suppliers can better anticipate demand and improve management of their inventory costs.

The assortment of use cases highlighted above depicts how utilizing Big Data Analytics can provide enormous value to enterprises. The value provided by Big Data Analytics is not only pertinent to certain industries but is relevant to virtually every industry as evidenced by the use cases outlined below which covers a range of different verticals.

## Financial Services

- Compliance and regulatory reporting.
- Risk analysis and management.
- Fraud detection and security analytics.
- Credit scoring and analysis.
- Trade surveillance.

## Government

- Fraud detection.
- Compliance and regulatory analysis.
- Energy consumption and carbon footprint management.

## Web & Digital Media Services

- Large-scale clickstream analytics.
- Ad targeting, analysis, and optimization.
- Social graph analysis and profile segmentation.
- Campaign management and loyalty programs.

## **Conclusion:**

It is a foregone conclusion that Big Data is beginning to take center stage given the explosion of data and the dire need of being able to glean useful business insights from them. Big Data Analytics provides the way for identifying useful pearls of wisdom from otherwise useless data. Big Data Analytics is becoming mission critical in the enterprises of the future.

**Appendix:**

<sup>1</sup>Information Management, "BIG DATA is Scaling BI and Analytics", Sept-Oct 2011