Fundamentals of Big Data and Smart Data Whitepaper

Table of Contents

1. What is Big Data?

Big data provides an innovative approach for transforming the way organizations are managed. This transformation is accomplished by using data more effectively and making data more accessible and transparent across departments, organizations, and industries. Using big data, your organization can have real-time insight into who did what, when, and where. The value created from this type of insight enables effective decision-making, real time monitoring of events, and the ability to find, analyze, and visualize data with your organization's tools of choice.

In the Gartner article titled, "How to Build Your Own Big Data Security Analytics Capability," Gartner defines big data as "high-volume, -velocity and –variety [sometimes called the 3Vs'] information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making…Big data is often associated with high volumes of data, distributed processing, the ability to link-together high-variety data (structured and unstructured), low-cost storage and computing, longer-term analysis, [non-relational] data stores, and other important characteristics."

The term "big data" is often misleading as big data projects often start out modestly with the variety and velocity of the data more important to project success than the volume. In other words, the size of the data is often less important than having high-quality "smart data" that organizational decision-makers can act upon. As a project progresses and succeeds, there is often a desire to grow the volume of data and identify more uses for that data. An example of a successful big data initiative is an organization that analyzes its data to determine the various types of data available, organizes the data to make needed data easily obtainable, and identifies the questions, trends, or metrics to which answers are needed.

2. What are the Big Data Success Factors?

Organizations that have successfully managed to implement a big data approach have the following three success factors in common: an analytics mindset, the required skillsets for implementing a big data approach, and the insight to seek-out lessons learned from organizations where a big data approach was successfully implemented. Each of these three success factors is described in greater detail in the following sections.

2.1. Analytics Mindset

An organization that is willing to use its own data, figure out all aspects of that data, and create new analytical approaches to investigate the data is implementing an analytics mindset. With this mindset, the organization takes a data-centric and data-driven approach and chooses to explore its data and discover what it needs to know from the data. This is as opposed to purchasing off-the-shelf solutions, relying on vendors to fix identified issues, and using out-of-box data such as security and validation rules as alternatives to solutions identified by exploring in-house data. While numerous applications are available to assist in discovering more about your organization's data, an inquisitive mind with a desire to explore data is the key to a successful analytics mindset.

An analytics' mindset is developed by using the tools to which you have access to explore your organization's data. Leverage the data available to you to make better organizational decisions and use the organizational tools available to you to discover new insights on the data or new areas of that data to explore. An analytics mindset is characterized by knowing that through the exploration of your organization's data, you have the ability to find new, factual answers to questions within your organization. As a first step in developing an analytics mindset, ask the following question, "What kinds of data are needed to answer the organization's existing questions?"

When answering this question, focus on how you can leverage existing data differently to obtain the answers as opposed to focusing on tools.

2.2. Required Skillsets

Organizations that have succeeded in implementing a big data approach have noted the importance of the following skillsets in achieving this success: statistics knowledge, domain knowledge, and technical knowledge in the areas of big data tools and analytics.

- ***Statistics Knowledge***. Statistics knowledge is an important skillset for understanding big data. Individuals with this skillset take into careful consideration aspects of the data such as the data collection process, the data modeling process, and sources of variation in the data. Individuals with this skillset are also known to make fact-based decisions as opposed to decisions based on personal experience or intuition.

- ***Domain Knowledge***. Domain knowledge is a necessity in terms of understanding which data, of all possible sources, is important to your organization. An individual with domain knowledge of the data under analysis has the unique ability to provide insight into that data. Without access to individuals with domain knowledge, your organization risks wasting a significant amount of time and effort in identifying the correct data to use in the correct manner for answering certain questions or resolving specific issues within your organization.

- ***Technical Knowledge***. Technical knowledge refers to having resources available who are subject matter experts (SMEs) in the areas of big data and analytics tools, infrastructure, and analytics. These SMEs focus on the operation of the organization's data and analytics tools and have experience working in a fast-paced, Agile environment characterized by the promotion of continuous improvement and the rapid delivery of solutions. Using an Agile environment to explore data allows technical SMES to get the answers they are seeking as quickly as possible. These SMEs have the knowledge and ability to install and run data analytics tools and are responsible for data storage, data access, and overall data management.

2.3. Lessons Learned from Big Data Success Stories

Assess your organization's needs to determine whether a big data approach is the correct approach. One way of accomplishing this is to seek out lessons learned from big data success stories. In addition, ask the following question: Is a more traditional approach that uses more traditional off-the-shelf tools a more appropriate approach for my organization? If your organization is not prepared or willing to transition to an analytics mindset and also invest in resources with the required skillsets for implementing a big data approach, your organization is not ready to implement a big data approach.

If you determine that your organization is both ready and committed to implementing a big data approach, consider the following recommendations:

- Review the skillsets currently available within your organization and identify where there are gaps.
- Analyze existing data and expand to new data types gradually after mastering the existing data and acquiring the necessary analytical skills.
- Focus on analysis before addressing big data collection.

- Learn from other organizations or other divisions within your organization that have successfully implemented a big data approach and mastered big data tools.
- Identify your organization's needs and determine whether the needed information is available within your organization's data.
- Recognize failures quickly and move on to explore the next question or problem.

3. What are Common Sources of Big Data?

Big data is the raw data that an organization collects and stores. There are numerous potential sources of big data that an organization can use to find answers to specific organizational questions, to identify trends in organizational data, to capture organizational metrics, and to make organizational decisions. While sources of big data are often internal to an organization, some organizations also look at sources of big data that are external to the organization.

Examples of common sources of big data are identified below:

- ***Transactional Data.*** Transactional data is data that has a time dimension (e.g. occurs at a specific time), such as: orders, invoices, and payments. When an organization uses a big data platform and analytics tools, the organization has the ability to obtain a greater understanding of transactions, such as the on-line behavior of users and the average time taken to complete on-line transactions. An organization can use this data to improve the user experience of its users.

- ***Log Data***. Log data is data that is automatically generated behind-the-scenes by a computer. Examples of log data include event logs and audit logs. Since log data typically produces large volumes of data, a big data platform is beneficial for collecting this data as it provides a scalable platform to collect the data and provides analytics tools that identify patterns in the log data that are otherwise not easily identified, such as issues related to application performance and user access.

- ***Business Application Data***. Business application data is data that is generated by business applications, including information that users input into business applications. For example: The VA My HealtheVet system allows patients to self-enter data such as demographics and emergency contacts, health care providers, health insurance, and treatment facilities. This data is an example of business application data.

- ***Data Storage Locations***. Data storage locations include document repositories, file storage locations, database servers, and big data tools such as: Hadoop, SQL, and NoSQL. Most data that resides in data storage locations is structured data, meaning the data has a defined format that allows for data analysis via a data analytics tool. An exception is many of the file formats associated with data that resides in document repositories, such as documents with the following extensions: .docx, .doc, .xlsx, .xls, .pptx. .ppt. The data that resides in these file formats is unstructured data that organizations often do not use. Since some business processes rely on paper or a combination of both paper and electronic data, an organization may find that there is considerable value in this unstructured data. This is especially true if the data is analyzed in conjunction with the electronic data.

- ***Internet of Things (IoT).*** IoT is "a computing concept that describes a future where every day physical objects will be connected to the Internet and be able to identify to other devices.[1]" In healthcare, an example of IoT is a hospital inpatient with multiple devices attached, each gathering different forms of data and all sharing the data with other devices. In theory, these devices are capable of determining whether a patient's heart rate drops below a certain range and responding accordingly without human intervention (i.e., change the medication infusion).[2]

- ***Archives.*** Archives refer to data collected in the past and retained for future reference such as medical records, statements, insurance records, legacy documents, and any original record between organizations or between an organization and their patients or clients. While the focus of big data is on real-time or near real-time data, organizations should consider that there may be data available today that is not important to the organization but that may be important to the organization in the future. An example is an organization that wants to compare data, such as metric data, over an extended period of time.

4. What are the Big Data Challenges?

While the benefits of big data are significant, organizations that implement a big data approach are also faced with several challenges that must be addressed in order to realize the full potential of big data. A list of some of the big data challenges are identified below.

- **Data Access.** Big data is raw data that often comes in formats such as dates, numbers, and dollar amounts. By itself, big data often has little value. The various sources of big data do not, on their own, typically provide the context needed by organizational decision makers to make decisions or answer questions. Big data must therefore be integrated or connected to other big data sources and data analytics tools to provide value to the organization.

- ***Data Preparation:*** In order to make big data useful, an organization must ensure that the data is in a format that is easily consumed by data analytics tools. Structured data is organized data that has a pre-defined data model, such as data that resides in database tables. Structured data is typically easily consumable by data analytics tools. Unstructured data; however, is data that does not have a pre-defined structure or data model. Examples of unstructured data include text files, audio files, and videos. When big data is also unstructured data or data that is not high-quality data, an organization must spend time preparing the data for use by data analytics tools.

- ***Data Availability and Performance***. When your organization relies on big data for real-time or near real-time data, down-time can cause challenges. When choosing big data and data analytics technologies, an organization needs to consider the downtown needed for these technologies and whether there is downtime built into these technologies. Likewise, the organization must consider the effect these technologies will have on the performance, or speed, of the data.

[1] What is the Internet of Things (IoT)? Definition from Techopedia (n.d.). Retrieved October 14, 2015 from http://www.techopedia.com/definition/28247/internet-of-things-iot

[2] Long, S. (2014, February 21). Integration vs. Interoperability: More than a Matter of Semantics. Retrieved May 13, 2015, from Caution-http://blog.capsuletech.com/integration-vs-interoperability-more-than-a-matter-of-semantics

- ***Data Reuse*.** Data is often collected and used for a single purpose. For example: The data in an electronic health record (EHR) is used by clinicians and health care providers to track the health information of a patient. Since reusing clinical data was not the intended purpose of the electronic health record, extracting data from EHRs for the purpose of reuse by data analytics tools is not always an easy task.

- ***Data Security***. To ensure that an organization's data remains secure, an organization must identify and follow-through with implementing techniques that secure data and prevent data breaches from occurring. This is especially important when an organization implements a new technology or technology platform that touches its data, the technology has few security mechanisms in place to protect the data, and there are specific organizational policies and procedures by which the organization must abide.

- ***Data Ownership.*** Implementing a big data approach often involves working across organizational functions and organizational divisions where there are different data owners. To address the challenge of having numerous data owners, organizations must discover new ways to collaborate with the various data owners throughout the organization.

5. How Does Big Data Become Smart Data?

Big data is raw data that often comes in formats such as dates, numbers, and dollar amounts. By itself, big data often has little value. When an organization extracts valuable data that is used by analytics and visualization tools and by organization decision makers to make decisions, the extracted data is known as smart data. Smart data addresses many of the big data challenges and is the necessary quantity and quality of specific types of data needed by an organization to provide value to the organization. The value to the organization is realized when trends, gaps, or outliers are seen in the data and organizational decision makers can make decisions based on the data.

One of the characteristics of an organization that successfully extracts smart data from its big data is the use of evidence-based decision making. Evidence-based decision making means that organizational decision makers are relying on high-quality data to make decisions. High-quality data is data that is correct, easily retrievable, fulfills its intended purpose, and is current and available in either real-time or near real-time. To make organizational decisions based on data, your organization must trust its data.

To achieve high-quality data, there is a need for strong data quality standards and a data governance body to enforce these standards. Data quality standards ensure that rules are in place regarding how your organization's data is collected and maintained. A data governance body ensures that resources with the necessary level of authority are in place to enforce compliance with the data quality standards. The overall goal of enforcing strong data quality standards and a data governance body is to ensure the availability and use of high-quality data within your organization. Using high-quality data will allow your organization to more easily extract smart data from its big data, which in turn makes the process of analyzing the smart data using data analytics tools much easier.

When an organization begins to analyze its data and is unknowingly collecting the wrong types of data or does not have the amount or quality of data expected, the organization will have incorrect analysis results without realizing it. Organizations must therefore regularly assess their data to make sure that the expected type, quantity, and quality of data is used for data analytics. Without high-quality data, meaningful data analytics is not possible and organizational decision makers do not have the ability to make effective decisions using data analytics tools.

6. How is Smart Data Used to Make Evidence-Based Decisions?

While the use of smart data is an essential aspect of making evidence-based decisions, the ability to visualize smart data is also very important. Visualizing smart data provides a way to tell a story with data and provides decision makers with a way to easily understand the data, make decisions based on the data, and get results from the data. Visualization tools allow decision makers to obtain a deeper insight into smart data and come to a more actionable analysis of the data by visually observing data trends and patterns. Common visualization tools include placing data in a format such as a chart, graph, infographic, or map. More advanced data visualization options include using smart data to tell a story via animations or videos.

Spreadsheets and dashboards are some additional examples of visualization tools commonly used by organizational decision makers. Spreadsheets are often used for performing data calculations and quickly analyzing and drawing conclusions from data on a one-time basis. Spreadsheets are not, however, an efficient reporting mechanism. Since spreadsheets do not automatically update, decision makers must manually update the data in spreadsheets to keep the information up-to-date. Web-based dashboards are another type of visualization tool that allows an organization to tell a story with its data by populating a dashboard with visuals. The content in a dashboard may be static or data-driven. With a data-driven dashboard, your organization can assist organizational decision makers by providing them access to real-time or near real-time data, such as metric data.

There are also data analytics tools and open-source big data analytics tools that allow an organization to visualize smart data and make decisions based on the data. Examples of big-data analytics tools include Hadoop, Spark, and Hive. These tools have the ability to collect, process, and analyze data. Some tools load data into a staging area before the data is loaded into a warehouse for analysis; however, an increasing number of big data vendors are creating tools that function as a central repository for an organization's incoming raw data. With this architecture, subsets of data are filtered for analysis in data warehouses or analytics' databases. As an alternative, the data can also be analyzed directly in big data tools with the help of various supporting technologies.

7. Conclusion

Big data is the raw data that organizations collect via data sources such as transactional data, log data, and business application data. By identifying the big data that provides value to an organization, enforcing the use of data standards to improve data quality, and using a governance body to validate the appropriate implementation of the data standards, your organization is transitioning its big data to smart data. Smart data provides your organization with the ability to view real time or near real-time data and make decisions based on this data. This can lead to greater efficiency within your organization and provide opportunities for more innovative uses of your organization's data.