BIG DATA: TERMS, DEFINITIONS, AND APPLICATIONS

Anuj Mediratta
Founder and Director
Ace Data Services Pvt. Ltd.

EMC²

Table of Contents

What is Big Data?

Big Data is a broad term used to refer to the huge volume of digital information generated by various businesses. This big data is not only generated by traditional information exchange and software, but also from sensors of various types embedded in a variety of environments; hospitals, metro stations, markets, and virtually every electrical device that produces data.

Big Data puts an inordinate focus on the issue of information volume. It exceeds the capacity of traditional data management technologies creating the need for new tools and technologies to handle the extremely large volume. It not only presents a challenge in storing large volumes of data but also the new capabilities of analyzing this huge volume of data.

Another way of defining Big Data is datasets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low latency. Big Data comes from databases, sensors, devices, audio/video, networks, log files, transactional applications, web, and social media mostly generated real-time and at a very large scale.

The term “Big Data” has dozens of definitions. Though largely focused on the volume of data, other V’s – i.e. variety and velocity – come into picture. IT leaders have begun to realize that that there are more than one challenge and dimension to data other than the new structures.

In the coming sections, we discuss the various new structures of data, the three V’s that form the basis of big data, and the new dimensions that have arisen as challenges while discussing the big data.

Data Structures

We must first understand the new types of data structures. Traditionally, we have been focused towards structured and unstructured data. Structured data is that which is contained in relational databases and spreadsheets. Structured data conforms to a database model having a fixed structure or format of capturing data. Database tools and additional reporting and analyzing tools have been used to help analyze this data and creating meaningful reports.

Unstructured data does not have a pre-defined data model nor is it organized in a pre-defined manner. It is typically text heavy and may contain data such as dates, numbers, and facts and include untagged data representing photos and graphic images. Word processing documents, presentations, and PDF files are prime examples of unstructured data.

New data structures that have come up are semi-structured data and quasi-structured data.

Semi-structured data is not the raw data and is not stored in a conventional database system. It is structured data but is not organized in a rational model like a table or an object-based graph. Semi-structured data contains tags or markers to separate semantic elements and enforce hierarchies of records and fields within the data. The entities belonging to the same class may have different attributes even though they are grouped together irrespective of the attributes' order. Markup languages like XML, email, and EDI are forms of semi-structured data. These support nested or hierarchical data simplifying the data models representing complex relationships between entities. These also support the lists of objects that simplify data models by avoiding messy translations of lists into a relational data model.

Quasi-structured data is more of a textual data with erratic data formats. It can be formatted with effort, tools, and time. This data type includes web clickstream data such as Google searches. Other examples are pasted texts which yields a network map based on similarity of language within the text, as well as proximity of words to each other within the text. However, it is not "tagged" the way YouTube and Flickr track content in images. Generally, to get untagged or image-based textual data, try an algorithm to analyze it and refine it based on the results you get.

It is believed that actual structured data is only 5% of the total data and therefore you need better ways to analyze the remaining 95%. Traditional database analysis or searching standard texts does not help complete the overall analysis.

Some people continue to classify data into structured and unstructured data only with semi-structured and quasi-structured as sub-types to unstructured to avoid confusion. More important is that organizations are realizing the benefits of analyzing data beyond databases and therefore moving beyond MIS reports.

Big Data is not just Volume

From the IT prospective, the first thing everyone wants to discuss when discussing data is the size of data. How large is it? How much physical storage do you need? When it comes to renaming this data as Big Data, the name itself implies that we are talking about really large volumes of data.

Data needs to be meaningful and should create meaningful results for the enterprise. Therefore, it is important to understand the characteristics of this Big Data so that ways and means of analyzing this data are better developed. It would also help in defining what result an enterprise should expect from this data.

Researchers defined the first characteristic model for understanding this by using the following three V's.

- **Volume**: The volume defines the size of the data that gets collected and stored. The concern is not only in terms of the storage required to store this but also the resources required for processing this huge amount of data irrespective of the source of the data and generate a real-time result from it. Data warehouse infrastructures and architects have been able to process this multi-terabyte of data at a reasonable cost for some large enterprises. With the emergence of highly scalable, low-cost storage and multi-core processing systems, analytics are becoming easier and more affordable for smaller enterprises while increasing the volumes to petabytes and exabytes for larger enterprises.

- **Velocity**: Data generation has changed from the traditional applications like invoicing or production where the data gets generated only during production hours and is restricted to how many invoices a day or how much production a day. New applications such as event-based alerting or flow-of-control monitoring require quick data processing. Enterprises now want to look at the results in a blink of an eye and see the impact of every new transaction. This has given rise to a new dimension in data analysis known as "streaming analysis".

- **Variety**: The volume is not coming from the structured data or database-based applications only. Datasets have a lot of new formats. Social media is one of the most important new varieties that have emerged. Enterprises want to capture how users are feeling about their products on social media and plan their future strategies accordingly. Other social media achievements include learning about spread of epidemic diseases and planning vaccination distribution accordingly. RFID applications, large quantities of PDFs, emails, recorded voice messages and videos, etc. add to the variety of data.

Complexity brought about by these three dimensions becomes the fourth dimension. Multiple factors will interact to make the challenge of data management more complicated.

For example, an increase in speed of generating new data pushes it to the velocity extreme. On the other side, large volumes get generated at that velocity and push it to the volume extreme along with it. Therefore, IT leaders cannot focus on only one extreme, they need to be prepared to handle all four at the same time and manage interaction between them. One way of handling this is to quickly analyze the data by bringing the application to the data as opposed to the traditional practice of taking data to the application. This can help handle multiple scenarios simultaneously.

New tools both in data capturing and applications help capture and process this large volume of data coming at a high velocity from a variety of new data sources. Platforms like Hadoop Distributed File System, MapReduce, and Hive enable engineering of new and innovative methods of data processing.

While the three V's described above help understand Big Data from a technical prospective, there is another 'V' model that helps understand the same from a business perspective. These include:

- **Variability**: Examine the datasets more closely. Establish if the contextual structure of the data stream is regular and dependable. Decide how the nature and context of text data content can be interpreted in a way that becomes meaningful for the required business analytic-ready models.

- **Veracity**: Verify if the data is suitable for its purpose and usable within the analytic model. Decide if it is trustworthy to analyze this data against a set of defined criteria. Capture the business procedures that enable the data to be profiled and validated. If

the problems are identified, undertake the right activities to remediate the data before any analysis is performed.

- **Value**: Identify the purpose, scenario, or business benefit that the analytic solution seeks to address. Thoroughly analyze what is the possible outcome and what is the desired outcome of the analysis. What actions need to be taken to get the desired results? Ensure that the analysis to be performed meets ethical considerations without reputation or compliance implications.

The three business oriented data characteristics mandate that a clear and shared understanding of the business context is established and communicated. The definition is used to frame the meaning and purpose of the data content. All three show the different aspects of the fitness for the purpose of datasets in question and how to deal with it to solve a business problem.

The new dimensions to Data

Big Data often starts the discussion about the new dimensions defined for data. These need to be handled in a different way than just handling Big Data. These new challenges are:

- **Real-time data**: This data is different from the traditional form of data that we store on our servers. It does not matter whether this falls under structured or unstructured data type. The key aspect is that this is about the “current data” not the old data. It enables situational awareness on what is happening now. Real-time data raises the issue of perishable and orphaned data which no longer has valid use cases but continues to be in use nonetheless.

- **Shared data**: This deals with the information that is shared across the organization. This includes sharing information between various applications and data sources. To share information efficiently, enterprises need to ensure that the data is consistent, usable, and extensible. The important aspect here is that sharing of information complicates the task of determining the authority of information.

- **Linked data**: This comes from the various data sources that have relationships with each other and maintain this context so as to be useful to humans and computers. Once a user links the data, relationship in that data persists from that point onwards.

- **High-fidelity data**: This data preserves the context, detail, relationships, and identities of important business information. This is largely done through the embedded metadata. High-fidelity data allows new meaning to be added without destroying the previous meaning of the data.

Traditional information management techniques assume cohesive control of both the storage processes and the integrity of the information. Then they link disparate sets via metadata instructions, which are executed in a type of application server. With big data, the process must move to the data, instead of moving the data from its stored location into a process and then back to write out the result. The only merit of the traditional approach is that it happens to be in place. However, it is yet to provide any real advantage and actually increases the number of hours required to maintain and modify it.

With big data coming in, traditional approaches will fail. Data and system architects have realized that immediate business needs drive the information architecture along one or more dimensions of data management sometimes excluding the other dimension. Hence, the moment any information leaves its original process, the excluded dimensions reassert themselves.

Analyzing the Data

Since IT came into existence, data always needed methods to analyze it and generate meaningful information from it. Large volumes of data being generated with reports being created out of them have been the traditional approach for businesses to operate and expand.

Database Analytics

The most basic type is database analysis wherein data is stored in the fixed row and column format of the tables and table in the database. Programmers would run queries on these tables to get the desired results and use another tool to present this data in a more logical manner. Reporting tools would help with graphs and so forth as well.

Database analytics continues to have relevance. Although new analytic dimensions are emerging, database analytics would never lose its charm as this is more towards analyzing the static data and getting reports out of it. Examples of database types are:

- **Relational**: This database type stores data in rows and columns. Parent-child relationship can be joined remotely on the server, providing speed over scale. This is the type of database where organizations usually start. These are good when you have highly structured data and you know what you will be storing. Production queries here will be very predictable. Database examples include Oracle, SQLite, and PostgreSQL, among others.

- **Document**: These databases store data in documents storing parent-child records in the same document. The server is aware of the fields stored within the document, can query on them, and return their properties selectively. These databases are good when your concept of record has relatively bounded growth and can store all related properties in the same document. Database examples include MongoDB, CouchDB, and BigCouch, among others.

- **Big-Table Inspired**: This database type stores the data into column-oriented stores inspired by Google's BigTable paper. It has tunable CAP parameters adjustable to provide either consistency or availability. Both of these adjustments are operationally intensive. These databases are good when you need consistency and write performance that scales past the capabilities of a single machine. Hbase is one such database type, which can help analyze data across 100 nodes in production.

- **Graph**: These databases use graph structures with nodes, edges, and properties to represent and store the data. There is no index adjacency. Every element contains a direct pointer to the adjacent element and no index lookups are necessary. These are good when you need to store collection of objects that lacked a fixed schema and linked together by relationships. Neo4j, OrientDB, Giraph, and Titan are examples of these.

- **NewSQL**: These databases are nearly like relational databases except that these offer high performance and scalability while preserving the traditional notions. Capable of high throughput online transaction processing requirements, these databases are the scalable version of a relational database that handles queries more efficiently. VoltDB and SQLfire are examples of this database type.

As needs increase, you find more scalable and high performance databases coming up to handle these. Big database analytics spread these across multiple nodes to eliminate compute requirement limitations as well. We will discuss the same in later sections of this article.

Real-time Analytics

Times have changed. Businesses cannot rely only on past performances to plan their future. They need to capture the current trends and the needs of the consumer today. Analytics have changed from analyzing past data to performing analytics on real-time data, generated not only from the structured in-house databases but also from non-structured data generated through social media and consumer behavior.

Tools are available today to collect data from these sources and put them together through what is called the process of “Data Conditioning” into databases to help analyze them. All of this data gets processed real time to produce near-instant results to help businesses serve their consumers better.

Many people ask if there is actually a need for real-time analysis. How would it help them in their business and is it worth the kind of investment it needs to get into real-time analysis?

You can analyze your business to see how you can use it for your business development. We can cite a few small examples of different streams that organizations have used to see the benefits. Let us start with healthcare – a small device worn on a waist belt injects desired quantities of insulin to a diabetic patient while monitoring the blood sugar level through the device.

Imagine you enter a hotel and the person at reception knows your name, booking details, and your preferences. While the second part is already in control through club memberships, the camera on the door takes your picture, searches for the record, and by the time you reach the reception, your details are on the desk.

Someone walking near the store gets a message on additional offers on their favorite product for the next 30 minutes. If you reject a product from the shelf, you get a message of additional offers on the product. A win-win for the store and the consumer.

Predictive Analytics

Organizations no longer ask for just analysis of their data. Traditional Business Intelligence (BI) tools have been doing that for a long time. What they are exploring is getting more useful insights from the data from visualization tools and predictive analysis to explore data in new ways and discover new patterns.

Predictive Analysis is the practice of extracting information from existing data to determine patterns and predict future outcomes and trends. It helps forecast what might happen in the future with an acceptable level of reliability and includes what-if scenarios and risk assessment.

Applied to business, predictive analysis models are used to analyze current data and historical facts to better understand customers, products, and partners and to identify potential risks and opportunities for a company. It uses a number of techniques, including data mining, statistical modeling, and machine learning to help analysts make business forecasts.

All businesses are run at a risk. Risk of the way business is managed. Every decision an organization takes impacts the risks an enterprise can withstand, i.e. the risk of customer defection, of not responding to an expensive glossy mailer, or offering a huge retention discount to a customer who was not leaving while missing out on a critical customer who leaves.

The data-driven means to compute risk of any type of negative outcome in general is predictive analysis. Insurance companies have used this very well, augmenting their practices by integrating predictive analysis in order to improve pricing and selection decisions.

The actuarial methods that enable an insurance company to conduct its core business perform the very same function as predictive models, rating customers by chance of positive or negative outcomes. Predictive modeling improves on standard actuarial methods by incorporating additional analytical automation and by generalizing to a broader set of customer variables.

Big Data Analytics

We have been discussing in detail about Big Data being not only large volume but also the variety of data sets and velocity at which it is generated. Now, we need to understand what Big Data Analytics means.

Simply put, it is the process by which we can examine these large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences, etc. to extract useful information for the business. It is the use of advanced analytic techniques against very large datasets that include real-time data collection from various structured/unstructured data sources.

Information without an insight is an unrealized resource. Conversely, analytics without a solid information foundation is likely to lead to poor decisions. Big Data analytics is the application of analytic capabilities on enormous, varied, or rapidly changing datasets. It is the application of analytic capabilities combined with the increased scope and context of big data, particularly when merged with traditional structured datasets.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions.

Big Data benefits to some Industries

Big Data concepts need to be implemented in many industries as they are collecting data from multiple sources and storing a great amount of this data. Collection and storage could initially be for compliance reasons. However, investing a bit more on top of it to get useful insight into the data is an added benefit.

Organizations now have an increased understanding of what Big Data is and what business benefits can be derived out of it. Big Data initiatives are becoming critical for organizations as they are not able to meet their business requirements with the traditional data sources, technologies, or practices. IT and business leaders are beginning to feel left behind in reaping the benefits of these new technologies.

However, most organizations who have even adopted Big Data are in the initial stages of planning their initiatives. Even those who have gone live tend to be function-specific and vary considerably in their business goals and business and technology requirements. This is making it challenging for them to choose the right approach and solution.

Large enterprises in many industries have started implementing their big data strategies and have started seeing benefits. Their peers in the same sector or even in another sector have also started learning from them and their experimentations to design a framework for themselves to implement their own strategies.

Big Data in Education

Various Big Data initiatives and applications have been helping large universities and education systems deliver better services to their students. There are different stages at which these initiatives help the education system.

In a pre-class initiative for example, big data helps match students with a learning opportunity that best fits their needs with the learning profile. Grants have been provided to education systems targeting creating systems that enable state and local education agencies to improve student learning and outcomes through data-based decisions. Such data may be used to assist schools in assigning students to appropriate learning environments and courses of study, and is especially helpful in understanding the needs of students who move between schools and school districts one or more times in a school year.

A post-class key activity is to achieve macro-scale improvements to areas such as institutional curriculum to help meet student and employer needs. This is something similar to when a company changes its business model to be more successful in a market it has chosen to target. The Predictive Analytics Reporting Framework Project represents over 600,000 student records and more than 3 million course records from only six institutions. Its goal was to look for the patterns on relationships that are not visible unless one can examine all institutional data from within a single multi-institutional database.

The initial analysis done at the end of eight months of the project generated meaningful results and was further expanded to involve another 10 institutes and drawing from a much larger database.

Another example came from the Italian government who collected data such as how quickly students get employed, their salaries, and the relevance of their education degrees in gaining employment. This can be used to gain critical insights on how the overall education system should be improved to deliver a better-educated generation to the market.

Big Data in Healthcare

Healthcare has large datasets. Even the smallest hospitals capture a lot of X-rays, scans, test records, etc. There are several use cases where hospitals can analyze the data and generate meaningful information. Data capturing is not an issue for them as they get real-time patient data, test results while the patient is undergoing care, and impact of medicines. These can be correlated with the patient's personal, environmental, and economic conditions data.

One example of healthcare utilizing this is real-time devices capturing patient's health parameters through a wearable device which captures key health parameters such as blood sugar levels, blood pressure, heart rate, etc. from time to time along with the body activity and environmental conditions, i.e. weather conditions. This data is automatically uploaded on the servers and doctors at remote locations can help the patient with medication without the need to travel to the hospital.

The same data sets can be used to find what kind of disease people are facing in different environmental conditions from time to time. Co-relating these with more data from social media can authenticate as well as expand the scope to achieve better results.

Hospitals can better understand the result of their medication on their patients as well as similar such results from other doctors and hospitals to make an informed decision on the next level of medication and treatment. Capturing experiences of multiple doctors and hospitals to make an informed decision adds a more scientific approach to treatment over and above the experience of the doctor.

Misconceptions about Big Data

Big Data has been around for some time now. However, not everyone knows everything about it. Several IT leaders have their concerns and doubts about Big Data technology.

Let us try and understand some of those and see if they really are of any concern:

- **80% of all data is unstructured:** This is one of the oldest misconceptions about data and data analytics. Given the variety of data sources, these appear true though these are not exactly true. There wouldn't be any patterns to discover in data if it had no structure. Use of non-relational databases like NoSQL and graph databases helps create the structures and the patterns for most data types.
- **Advanced analytics is just an advanced version of Normal Analytics:** Many believe they have mastered database analytics and just need next step to move onto advanced analytics. Normal analytics are mainly static reports from the static databases. Advanced analytics give us the power to reach intelligent conclusions and solve real business problems from analyzing the data.
- **Embedded analytics solves all problems:** Embedded analytics are the standard set of tools or reports embedded over the data and data sets. Like normal analytics, these also only deliver reports and do not really analyze data on patterns and results to provide meaningful outcomes.
- **Improved tools will replace the Data Scientist:** Regardless of the type of tool or advancement of the tools, you would need data scientists/analysts to use these tools to perform the analysis and get dynamic reports. In any case, there is a shortage of data scientists so they would always be required.
- **Data Scientists need high-level education:** Data Scientists need an analytical, logical mind to define the rules and relationships between data sources to get positive outcomes. We believe education solely does not help if application of mind and logic is not done properly. It is more important to be more logical and understand the business needs.

- **We can predict everything with Big Data:** While we can use big data to form patterns and predict many things, big data cannot predict everything. Hospitals can analyze which kind of people are at higher risk of heart ailments so that precautions can be taken but many things in more complex domains such as law and politics cannot be predicted.
- **Big Data isn't biased:** Data is always biased regardless of the volume or data source. Data is a result of certain measurements and was collected with some purpose. You need to approach it with care and be careful that you have collected the dataset from the representative of the group you are studying or else you get the wrong data.

Building a Big Data strategy

It is important for an organization to form the basic strategy of their big data initiative. The organization and the stakeholders of the project associated with it in any form need to be well informed about the strategy.

The analysts need to work in a very transparent way and be able to show what tools and data sources they are going to use and the conclusion expected from it. They need to demonstrate all possible options to the management with a clear analysis and cost benefit. Another key aspect is privacy and security of the information especially when you are handling personal data. The most important thing is to respect the opinions of others as they are closer to their data.

Let us form some basic principles of our analysis strategy:

- **Get the basics right and simple:**
  i. Never break the law and regulations that your enterprise needs to follow.
  ii. Be professional and use up-to-date analytical techniques. Be prepared to reject every hypothesis.
  iii. Respect the point of view of other professionals.
  iv. Take care of the security.
- **Information Governance needs to grow:** Next to technological solutions, organizational measures help to prevent ethical issues. Like with all staff handling sensitive issues, data scientists should regularly sign a legally binding nondisclosure agreement.

- **Connect business to consumer value:** This is an essential aspect as all we are doing is ensuring better value being delivered to our consumers. We cannot make wrong decisions or analyze wrong data sets that do not deliver appropriate value to our consumers.
- **Be sensitive to cultures and values:** These will vary for different organizations or industries. You need to understand these very well as they form the basic understandings and relationships. You should know what the organizations want from you. What are the stakeholders expecting from your analysis? What are the expected business practices in your industry? In a multicultural environment, on what analytical initiatives are we prepared to differentiate and adapt to local opportunities? Where are we not prepared to differentiate and why?
- **Understand the origin of the data:** Data is collected for a specific purpose and with a measurement instrument. In essence, from the time it is collected, data represents a point of view. Don't collect the data for the sake of its availability and potential opportunity to use it in the future. Collect it for your current purpose.
- **There is nothing objective in analytics:** Many people want analysts to aim at truthfulness in their work. However, since the analysis depends on the primary data collected, an analyst can only ensure that the most appropriate data for the analysis has been collected.

Business Essentials of Big Data Strategy

It is important to understand that while we are discussing data, Big Data initiatives are not about investing more in IT only. Big Data initiatives differ by businesses and need IT as a support function underlying them. Big Data initiatives can yield huge positive results by enabling unprecedented agility, generating goodwill, and boosting product and process innovations.

Big Data initiatives are about change in the approach to how we look at our business. It moves away from traditional BI and predicts growth to generate real business analysis. Even data collection has moved from punching invoices to generating data from sensors and gathering data from social media to see how products and services are being accepted in the market.

Businesses need to be prepared to get a different result that they never thought about for their products and market acceptance. Decision-making is no longer intuition-based; it is now more real data-based and data samples have a huge variety so the possibility of better decision-making increases.

Big Data initiatives focus more on acquiring, integrating, and preparing information. This is another significant shift from the traditional data's functionality, which is more or less straight forward as identifying correlations, anomalies, or patterns. This shift in focus has brought in the enterprise architecture, project management, and role definition.

From the IT perspective, Big Data differs greatly from traditional technologies even if they use state-of-the-art hardware database management systems and analytics capabilities. Big Data initiatives need an entirely different environment prepared to collect and process an enormous volume of data collected from varying data sources at real-time velocity.

This brings in the financial aspect of the strategy as it requires huge investments in the strategy. Thus, it becomes very important for the organization to ask questions such as "What value can we generate out of this data?" "How much can we accommodate, administer, and apply it?" Business initiatives that are most successful are those which measure the possible business outcome and map it with their investments.

Big Data is not a static technology that an organization can purchase and deploy. It is a strategy that the organization needs to form for itself. It is an idea that an organization needs to develop and apply for its betterment. Organizations need to ask relevant questions before they formulate the strategy. These need to be specific yet open-ended, should relate to the business process and, most importantly, should focus on optimizing or innovating, not merely informing.

Organizations need to consider changes relative to other indicators and sensors, leverage internal and external inputs, and be forward looking. The strategy needs to consider various business scenarios and data sources. It needs to move from a "prove it" to a "do it" strategy. The most important thing to realize is that it should move on from just comparing yourself to differentiating yourself.

Organizations need to be prepared for a radical change in their processes. Remember, the results are going to come based on the data sources and you are expected to involve social media, sentiment analysis, etc. also and these could change your overall approach to your business.

Big Data can help product management, sales, and marketing teams in various ways:

- Identify new markets.
- Identify new feature needs for the product.
- Identify opportunities for completely new offerings.
- Personalizing and mass communication.
- Aggregating, packaging, and selling information products.

New markets and geographies can be determined by analyzing available public or commercial data for gaps among competitors or between distinct markets. Opportunities for new features can be discovered through analysis of aggregated customer touch points or broader social media. New offerings can be created by performing predictive analysis against combinations of market and feature data or by listening to social media influencers.

Finally, enterprises are starting to discover the economic value of productizing the data they generate and collect.

Variable Data Sources

Great ideas on Big Data come from understanding the range of data sources available and what questions can be answered if they are integrated and correlated. Data sources are the most important aspect of the strategy; results are affected by the choice of data and data sources. These data sources could be classified as:

- **Operational Data:** This constitutes online transaction processing and/or analytical processing databases. These are the traditional databases maintaining customer, employee, vendor details, etc. They can even collect sensor-based data like smart meters, Internet-connected devices, etc.
- **Dark Data:** This constitutes the old data an organization has collected over the period of running their business. This could be in archives or least accessed kind of data. Extracting usable data out of this data through parsing, tagging, linking, or structuring is considered the greatest opportunity by most businesses among all types of data.
- **Commercial Data:** This constitutes the data that survey or research organizations would have collected over a period of time. For example, surveying for a new technology amongst the CIO community and then collecting their information which is later shared/sold to other prospects.

- **Public Data:** Many governments have begun opening their data coffers. These are done to support economic development and health, welfare, and citizen services at various stages of implementation. These have significant mercantile value to understand local and global market conditions, population trends, weather, etc.
- **Social Media Data:** Social media participation is experiencing phenomenal growth. Facebook and Linkedin are updating their fast-growing, invaluable source of data about preferences, trends, attitudes, behavior, products, and companies. Posts, trends, and even usage patterns are increasingly used to identify and forecast target customers and segments, market opportunities, competitive threats, business risks, and even selecting ideal employment candidates.

More about investments

Big Data doesn't dramatically alter the economics of acquiring, administering, and applying information assets but it does amplify it. Organizations can no longer ignore the need to balance these information supply chain costs with the tangible value derived out of them. CIOs and CFOs need to align with how information asset costs and benefits are measured.

An important aspect in balancing Big Data benefits to outweigh its cost is ensuring the data serves multiple business purposes. Compiling, hosting, and processing petabytes of data for a single business process rarely makes sound financial sense or good use of scarce skillsets.

From an accounting perspective, data acquisition costs are explicit and could be capitalized as opposed to being considered as an operational expense. The cutting-edge integration, management, storage, processing, and analytics technologies often demanded by big data initiatives need special investment considerations.

Cloud-based hosting, novel predictive analytics products, and NoSQL database management systems like Hadoop can even offer economic advantages over traditional DBMS, on-premises, and enterprise BI solutions. Organizations can opt for starting on a cloud hosting model and start building their own environments once they begin harnessing results of their initial cloud hosting initiatives.

Key Strategic Considerations

The following considerations should be kept in mind while finalizing Big Data initiatives:

- **Is this the right time for a Big Data initiative for me?**

  Big Data is more about volume, velocity, and variety. However, there is no defined threshold of either type that defines that you need it now. It is a strategic decision for an organization to decide when they need a big data initiative and what kind of value are they looking to derive out of it.

- **Are others in my industry working with Big Data?**
  While this may not be relevant for organizations who want to pioneer a Big Data project in their industry, some prefer to be followers and learn from the experiences of others.
- **What are those in my industry working with Big Data achieving with it?**
  This would be important and interesting to know. Rather than trying to compete with leaders in the industry, you should get their insights on the benefits and emulate what would help you in your initiative.

- **What range of data sources is going to be useful for me?**
  Choosing the right data sources is the key for any Big Data initiative. There are two important aspects: what data source is more relevant, i.e. a consumer goods initiative might find more use in social media and this would also help define the cost for your initiative.
- **What is the value of Big Data project for me?**
  As for any other project, you need to define your key goals and the value you are looking for in the project. As discussed earlier, analysis results may not be of your choice or anticipation, so you need to be prepared for it.
- **What skills do we need for Big Data?**
  A lot of focus should be given to the skills for the Big Data initiative. Organizations do need Data Scientists and programmers who can help configure, administer, and manage the overall project. Also required is the experience in deploying and tuning multiple elements of operating systems, networks, and storage.
- **How do regulatory and security concerns impact the data we use?**
  This is the most complex issue. In many situations, data scientists, architects, and even business owners are unsure about which data can or should be used and for what purpose. Multinationals need to be additionally sure about what kind of data

privacy policies and cultural ethics some of the countries follow so that their branches worldwide are able to handle this better.

IT essentials of Big Data Initiatives

Many traditional and state-of-the-art technologies were not designed to handle large transaction processing especially for today's and tomorrow's level of data volume, velocity, and variety. Even as the data grows exponentially along these three dimensions, investments required for scaling technologies processors, storage, DBMSs, and analytics to perform sufficiently can grow even faster. There are a variety of ways to handle these intractable economics. These include:

a. DBMS advances in loading, indexing, and parallelism
b. Grid computing
c. NoSQL databases
d. Distributed File systems
e. In-memory databases
f. Map-reduce processing
g. Usage-driven tiered storage
h. Cloud-based data and processing

Big Data introduces risks while handling the data. Data sources may include personal, sensitive, or proprietary information that can be prone to mishandling and misuse. Even if the individual data sources may not contain explicit information, integration of multiple sources could expose corporate secrets or identify individuals. This risk can be especially perilous when information is shared outside the organization with business partners, suppliers, trade organizations, or government.

As business benefits increase, so do the challenges in developing and implementing analytic approaches. Part of this challenge is sourcing, integrating, and analytically processing data. For this reason, Big Data's value for hindsight-only analytics is limited and economically indefensible. Conversely, Big Data's utility and value proposition is much greater for higher order analytics that provide deeper understanding, broader relevancy, and farther visibility.

Organizations also need to assemble the necessary skills for their initiatives. These skills include data integration and preparation, business and analytic modelling, collaboration, communication, and creativity. These skills are in extremely short supply. Therefore, organizations need to develop a three-pronged strategy to develop them:

1. Aggressively recruit from universities, competitors, and other initiatives
2. Develop skills internally through incentives and trainings
3. Leverage available consulting talent to fill gaps and mentor

Structures for the IT organizations also need a change. Big Data initiatives should be driven by a team that constitutes both skilled IT people as well as people who understand the business and the relevance of the data and its reports. As the Big Data solutions become ready for operations, IT should be in a position to take over the management of the infrastructure, architecture, and application components. New roles are forming in the IT organization for handling these:

- Chief Data Officer
- Chief Analytics Officer

Tools for Big Data Analytics

Below are some of the tools that can be used to perform various Big Data analytics-related functions to derive the desired results:

- **MapReduce**: A software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or standalone computers. The framework is divided into two parts:
    a. Map: A function that parcels out work to different nodes in the distributed cluster.
    b. Reduce: A function that collates the work and resolves the results into a single value.

    The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

MapReduce is important because it allows ordinary developers to use MapReduce library routines to create parallel programs without having to worry about programming for intra-cluster communication, task monitoring, or failure handling.

- **Hadoop**: Hadoop is an open source framework for distributed storage and processing of large sets of data on commodity hardware. Hadoop enables businesses to quickly gain insight from massive amounts of structured and unstructured data. It is designed to scale from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer. Hadoop enables a computing solution that is:
    a. **Scalable:** New nodes can be added as needed without changing the data formats, how data is loaded, how jobs are written, or the applications on the top.
    b. **Cost-effective:** Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.
    c. **Flexible:** Hadoop is schema-less and can absorb any type of data from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.
    d. **Fault tolerant:** When a node is lost, the system redirects work to another location of the data and continues processing without missing a fright beat.
- **Hive**: Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time, this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

- **Pig**: Pig is a platform for analysing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turn enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of MapReduce programs, for which large-scale parallel implementations already exist (e.g. the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

a. **Ease of programming:** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
b. **Optimization opportunities:** The way tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
c. **Extensibility:** Users can create their own functions to do special-purpose processing.

- **MADlib**: MADlib is an open-source library for scalable in-database analytics that can help improve data analysis efficiency and accuracy. It provides data parallel implementations of mathematical, statistical, and machine-learning methods for structured and unstructured data. These SQL-based algorithms for machine learning, data mining, and statistics run at speed and scale.

These are some of the key tools that can be used while developing a big data initiative. The choice would depend on the objective of the initiative and to some extent on the skillsets of the team involved. More can be discussed for each of these tools. I am keeping it short here as the idea of this article is not to focus on too much detail on the tools only.

Case Studies

We have discussed some details of how Big Data initiatives have been used by leading education institutes and healthcare service providers. Big Data can help solve problems and improvise operations, production, and overall business across all industries. At this stage, many of them could be just great ideas but they would get implemented soon to help derive results from them. Let us examine a few more examples in brief from some more industries:

- **Retailing**: Enhanced insights and understanding of customer likes and dislikes, influences, and behaviors can lead to increased sales, stock performance, and context-awareness in-store. One of the examples includes a situation when you are in a store and leave an article after looking at its price. The store intelligently recognizes you and the product and quickly sends you a text message for a special offer on the product. This not only has the potential of influencing you to buy the product and therefore increasing sales for the store but also feeling special with that personalized service happening automatically.

- **Online Social Media**: You are likely familiar with incidents on Linkedin where it shows the names of people others visited after you visited the same profile. This is a small initiative which has the potential to help you link with more known acquaintances faster than searching for them.

- **Online Stores**: What about getting offers on books of your favorite authors while purchasing clothes for yourself on an online store? For your convenience, the store adds what you like to purchase the most. Again, potentially resulting in better sales and higher customer satisfaction for repeat visits.

- **Public Sector**: Detailed real-time information regarding traffic flows, vehicle locations, and resource use supports an organization's service delivery and efficiency across a wide range of public sectors.

- **Healthcare**: Understanding lifestyle patterns and environment of a region could help Data Scientists predict a possible epidemic or disease spreading in a given region. This could also help authorities prepare better when the disaster strikes and distribute vaccines accordingly.

Conclusion

Big Data can be a complex concept for beginners. However, most organizations that have initiated this journey have realized benefits very quickly. Be aware that this does require building skill sets and investing on the resources. Therefore, it is critical to ensure that the investment vs value proposition is calculated properly while designing the initiative.

Use Big Data only when you feel there is a chance to get a more meaningful insight for your business. It requires “big thinking”, so when developing your ideas, keep in mind that Big Data analytics differs from traditional data mining or MIS reporting.

Focus on your real needs and choose your data sources, tools, and hardware accordingly. Combine big data with cloud computing to handle the practical requirements of scaling up and down quickly and get quick and relatively less expensive test environments.

Start with your prime objective and then start using the data to get more insight as it develops further.

Appendix

1. "Introduction to Big Data and Hadoop", http://pendhari.com/introduction-to-big-data-and-hadoop/

2. "Apache Hadoop Training", http://www.smc-sol.com/training/hadoop_training.html