**Big Data and Its Importance**

Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises. The greatest part about this phenomenon is the rapid pace of innovation

and change; where we are today is not where we’ll be in just two years and definitely not where we’ll be in a decade.

Just think about all the great stories you will tell your grandchildren about the early days of the twenty-first century, when the Age of Big Data Analytics was in its infancy.

This new age didn’t suddenly emerge. It’s not an overnight phenomenon. It’s been coming for a while. It has many deep roots and many branches. In fact, if you speak with most data industry veterans, Big Data has been around for decades for firms that have been handling tons of transactional data over the years—even dating back to the mainframe era. The reasons for this new age are varied and complex, so let’s reduce them to a handful that will be easy to remember in case someone corners you at a cocktail party and demands a quick explanation of what’s really going on. Here’s our standard answer in three parts:

1. **Computing perfect storm.** Big Data analytics are the natural result of four major global trends: Moore’s Law (which basically says that technology always gets cheaper), mobile computing (that smart phone or mobile tablet in your hand), social networking (Facebook, Foursquare, Pinterest, etc.), and cloud computing (you don’t even have to own hardware or software anymore; you can rent or lease someone else’s).
2. **Data perfect storm.** Volumes of transactional data have been around for decades for most big firms, but the flood gates have now opened with more *volume*, and the *velocity* and *variety—*the three Vs—of data that has arrived in unprecedented ways. This perfect storm of the three Vs makes it extremely complex and cumbersome with the cur- rent data management and analytics technology and practices.
3. **Convergence perfect storm.** Another perfect storm is happening, too. Traditional data management and analytics software and hard- ware technologies, open-source technology, and commodity hardware are merging to create new alternatives for IT and business executives to address Big Data analytics.

## The Sources of Big Data

### **Black Box Data**

This is the data generated by airplanes, including jets and helicopters. Black box data includes flight crew voices, microphone recordings, and aircraft performance information.

### **Social Media Data**

This is data developed by such social media sites as Twitter, Facebook, Instagram, Pinterest, and Google+.

### **Stock Exchange Data**

This is data from stock exchanges about the share selling and buying decisions made by customers.

### **Power Grid Data**

This is data from power grids. It holds information on particular nodes, such as usage information.

### **Transport Data**

This includes possible capacity, vehicle model, availability, and distance covered by a vehicle.

### **Search Engine Data**

This is one of the most significant sources of big data. Search engines have vast databases where they get their data.

Additionally, Bernard Marr, a Big Data and Analytics expert, has come up with his brilliant list of 20 Big Data sources that are freely available to everybody on the web.

**Five V’s of Big Data**

**1. Volume:**

* The name ‘Big Data’ itself is related to a size which is enormous.
* Volume is a huge amount of data.
* To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
* Hence while dealing with Big Data it is necessary to consider a characteristic ‘Volume’.
* *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

**2. Velocity:**

* Velocity refers to the high speed of accumulation of data.
* In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
* There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
* Sampling data can help in dealing with the issue like ‘velocity’.
* *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

**3. Variety:**

* It refers to nature of data that is structured, semi-structured and unstructured data.
* It also refers to heterogeneous sources.
* Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.
  + **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
  + **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.
  + **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn’t fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can’t be stored in the form of rows and columns.

**4. Veracity:**

* It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
* Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
* *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

**5. Value:**

* After having the 4 V’s into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
* Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s.

**Drivers for Big Data**

Big Data emerged in the last decade from a combination of business needs and technology innovations. A number of companies that have Big Data at the core of their strategy have become very successful at the beginning of the 21st century. Famous examples include Apple, Amazon, Facebook and Netflix.

A number of business drivers are at the core of this success and explain why Big Data has quickly risen to become one of the most coveted topics in the industry. Six main business drivers can be identified

**1. The digitization of society**

Big Data is largely consumer driven and consumer oriented. Most of the data in the world is generated by consumers, who are nowadays ‘always-on’. Most people now spend 4-6 hours per day consuming and generating data through a variety of devices and (social) applications. With every click, swipe or message, new data is created in a database somewhere around the world. Because everyone now has a smartphone in their pocket, the data creation sums to incomprehensible amounts. Some studies estimate that 60% of data was generated within the last two years, which is a good indication of the rate with which society has digitized.

**2. The plummeting of technology costs**

Technology related to collecting and processing massive quantities of diverse (high variety) data has become increasingly more affordable. The costs of data storage and processors keep declining, making it possible for small businesses and individuals to become involved with Big Data. For storage capacity, the often-cited Moore’s Law still holds that the storage density (and therefore capacity) still doubles every two years. The plummeting of technology costs has been depicted in the figure below. Besides the plummeting of the storage costs, a second key contributing factor to the affordability of Big Data has been the development of open source Big Data software frameworks. The most popular software framework (nowadays considered the standard for Big Data) is Apache Hadoop for distributed storage and processing. Due to the high availability of these software frameworks in open sources, it has become increasingly inexpensive to start Big Data projects in organizations.

**3. Connectivity through cloud computing**

Cloud computing environments (where data is remotely stored in distributed storage systems) have made it possible to quickly scale up or scale down IT infrastructure and facilitate a pay-as-you-go model. This means that organizations that want to process massive quantities of data (and thus have large storage and processing requirements) do not have to invest in large quantities of IT infrastructure. Instead, they can license the storage and processing capacity they need and only pay for the amounts they actually used. As a result, most of Big Data solutions leverage the possibilities of cloud computing to deliver their solutions to enterprises.

**4. Increased knowledge about data science**

In the last decade, the term data science and data scientist have become tremendously popular. In October 2012, Harvard Business Review called the data scientist “sexiest job of the 21st century” and many other publications have featured this new job role in recent years. The demand for data scientist (and similar job titles) has increased tremendously and many people have actively become engaged in the domain of data science. As a result, the knowledge and education about data science has greatly professionalized and more information becomes available every day. While statistics and data analysis mostly remained an academic field previously, it is quickly becoming a popular subject among students and the working population.

**5. Social media applications**

Everyone understands the impact that social media has on daily life. However, in the study of Big Data, social media plays a role of paramount importance. Not only because of the sheer volume of data that is produced everyday through platforms such as Twitter, Facebook, LinkedIn and Instagram, but also because social media provides nearly real-time data about human behaviour.

Social media data provides insights into the behaviours, preferences and opinions of ‘the public’ on a scale that has never been known before. Due to this, it is immensely valuable to anyone who is able to derive meaning from these large quantities of data. Social media data can be used to identify customer preferences for product development, target new customers for future purchases, or even target potential voters in elections. Social media data might even be considered one of the most important business drivers of Big Data.

**6. The upcoming internet of things (IoT)**

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and network connectivity which enables these objects to connect and exchange data. It is increasingly gaining popularity as consumer goods providers start including ‘smart’ sensors in household appliances. Whereas the average household in 2010 had around 10 devices that connected to the internet, this number is expected to rise to 50 per household by 2020. Examples of these devices include thermostats, smoke detectors, televisions, audio systems and even smart refrigerators.

## ****Big Data Applications****

The primary goal of Big Data applications is to help companies make more informative business decisions by analyzing large volumes of data. It could include web server logs, Internet click stream data, social media content and activity reports, text from customer emails, mobile phone call details and machine data captured by multiple sensors.

Organisations from different domain are investing in Big Data applications, for examining large data sets to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. In this blog we will we be covering:

## ****Big Data Applications: Healthcare****

The level of data generated within healthcare systems is not trivial. Traditionally, the health care industry lagged in using Big Data, because of limited ability to standardize and consolidate data.

But now Big data analytics have improved healthcare by providing personalized medicine and prescriptive analytics. Researchers are mining the data to see what treatments are more effective for particular conditions, identify patterns related to drug side effects, and gains other important information that can help patients and reduce costs.

With the added adoption of mHealth, eHealth and wearable technologies the volume of data is increasing at an exponential rate. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of data.

## ****Big Data Applications: Manufacturing****

Predictive manufacturing provides near-zero downtime and transparency. It requires an enormous amount of data and advanced prediction tools for a systematic process of data into useful information.

Major benefits of using Big Data applications in manufacturing industry are:

* Product quality and defects tracking
* Supply planning
* Manufacturing process defect tracking
* Output forecasting
* Increasing energy efficiency
* Testing and simulation of new manufacturing processes
* Support for mass-customization of manufacturing

Big Data provides actionable points of information about millions of individuals. Now, publishing environments are tailoring advertisements and content to appeal consumers. These insights are gathered through various data-mining activities. Big Data applications benefits media and entertainment industry by:

* Predicting what the audience wants
* Scheduling optimization
* Increasing acquisition and retention
* Ad targeting
* Content monetization and new product development

## ****Big Data Applications: Internet of Things (IoT)****

Data extracted from [***IoT***](https://www.edureka.co/blog/iot-tutorial/) devices provides a mapping of device inter-connectivity. Such mappings have been used by various companies and governments to increase efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data is used in medical and manufacturing contexts. 

## ****Big Data Applications: Government****

The use and adoption of Big Data within governmental processes allows efficiencies in terms of cost, productivity, and innovation. In government use cases, the same data sets are often applied across multiple applications & it requires multiple departments to work in collaboration.

Since Government majorly acts in all the domains, thus it plays an important role in innovating Big Data applications in each and every domain. Let me address some of the major areas:

### **Cyber security & Intelligence**

### **Crime Prediction and Prevention**

### **Pharmaceutical Drug Evaluation**

### **Scientific Research**

### **Weather Forecasting**

### **Tax Compliance**

### **Traffic Optimization**

**Bid Data Technologies**

Big data technology and Hadoop is a big buzzword as it might sound. As there has been a huge increase in the data and information domain from every industry and domain, it becomes very important to establish and introduce an efficient technique that takes care of all the needs and requirements of clients and big industries which are responsible for data generation. Earlier the data was being handled by normal programming languages and simple structured query language but now these systems and tools don’t seem to do much in case of big data.

Big data technology is defined as the technology and a software utility that is designed for analysis, processing, and extraction of the information from a large set of extremely complex structures and large data sets which is very difficult for traditional systems to deal with. Big data technology is used to handle both real-time and batch related data. Machine learning has become a very critical component of everyday lives and every industry and therefore managing data through big data becomes very important.

### Types of Big Data Technologies

Before starting with the list of technologies let us first see the broad classification of all these technologies.

They can mainly be classified into 4 domains.

1. Data storage
2. Analytics
3. Data mining
4. Visualization

**Let us first cover all the technologies which come under the storage umbrella.**

1. **Hadoop**: When it comes to big data, Hadoop is the first technology that comes into play. This is based on map-reduce architecture and helps in the processing of batch-related jobs and process batch information. It was designed to store and process the data in a distributed data processing environment along with commodity hardware and a simple programming execution model. It can be used to store and analyze the data present in various different machines with high storage, speed, and low costs. This forms one of the main core components of big data technology which was developed by the Apache software foundation in the year 2011 and is written in Java.
2. **MongoDB**: Another very essential and core component of big data technology in terms of storage is the MongoDB NoSQL database. It is a NoSQL database which means that the relational properties and other RDBMS-related properties do not apply to it. It is different from traditional RDBMS databases which make use of structured query language. It makes use of schema documents and the structure of data storage is also different and therefore they are helpful in holding a large amount of data. It is a cross-platform document-oriented design and database program that makes use of JSON like documents along with schema. This becomes a very useful use-case of [operational data stores](https://www.educba.com/operational-data-stores/) in the majority of financial institutions and thereby working to replace the traditional mainframes. MongoDB handles flexibility and also a wide variety of data types at high volumes and among distributed architectures.
3. It is useful in accessing data through remote Hadoop clusters by making use of virtual indexes and also makes use of Splunk search processing language which can be used for the analysis of data. The hunk can be used to report and visualize huge amounts of data from the Hadoop and NoSQL databases and sources. It was developed by team Splunk in the year 2013 which was written in Java.
4. Cassandra forms a top choice among the list of [popular NoSQL databases](https://www.educba.com/what-is-nosql-database/) which is a free and an open-source database, which is distributed and has a wide columnar storage and can efficiently handle data on large commodity clusters i.e. it is used to provide high availability along with no single failure point. Among the list of main features includes the ones like distributed nature, scalability, fault-tolerant mechanism, MapReduce support, tunable consistency, query language property, supports multi-data center replication, and eventual consistency.

**Next lets us talk about the different fields of big data technology i.e. Data Mining.**

1. **Presto**: It is a popular open-source and a SQL-based distributed query Engine which is used for running interactive queries against the data sources of every scale and the size ranges from Gigabytes to Petabytes. With its help, we can query data in Cassandra, Hive, proprietary data stores, and relational database storage systems. This is a java based query engine that was developed by the Apache foundation in the year 2013. A few sets of companies that are making good use of the Presto tool are Netflix, Airbnb, Checkr, Repro, and Facebook.
2. **ElasticSearch**: This is a very important tool today when it comes to searching. This forms an essential component of the ELK stack i.e. the elastic search, Logstash, and Kibana. ElasticSearch is a Lucene library-based search engine which is similar to Solr and is used to provide a purely distributed, full-text search engine which is multi-tenant capable. It has a list of schema-free JSON documents and an HTTP web interface. It is written in the language JAVA and is developed by Elastic company in the company 2012. The names of a few companies which make use of elasticsearch are: LinkedIn, StackOverflow, Netflix, Facebook, Google, Accenture, etc.

**Now let us read about all those big data technologies which are a part of Data analytics:**

1. **Apache Kafka**: Known for its publish-subscribe or pub-sub as it is popularly known as, is a direct messaging, asynchronous messaging broker system that is used to ingest and perform data processing on real-time streaming data. It also provides a provision of the retention period and the data can be channelized by means of a producer-consumer mechanism. It is one of the most popular streaming platforms which is very similar to the enterprise messaging system or a messaging queue. Kafka has launched many enhancements to date and one major kind is that of Kafka confluent which provides an additional level of properties to Kafka such as Schema registry, Ktables, KSql, etc. It was developed by the Apache Software community in the year of 2011 and is written in Java. The companies which are making use of this technology include Twitter, Spotify, Netflix, Linkedin, Yahoo, etc.
2. **Splunk**: Splunk is used to capture, correlate and index real-time streaming data from a searchable repository from where it can generate reports, graphs, dashboards, alerts and data visualizations. It is also used for security, compliance and application management and also for web analytics, generating business insights and business analysis. It was developed by Splunk in Python, XML, Ajax.
3. **Apache Spark**: Now comes the most critical and the most awaited technology in the domain of Big data technologies i.e. Apache Spark. It is possibly among the ones which are topmost in demand today and makes use of Java, Scala or Python for its processing. This is used to process and handle the real-time streaming data by making use of Spark Streaming which uses batching and windowing operations to make that happen. Spark SQL is used to create data frames, datasets on top of RDDs and thereby providing a good flavor of transformations and actions which form an integral component of Apache Spark Core. Other components such as Spark Mllib, R and graphX are also useful in the case of analysis and doing [machine learning and data science](https://www.educba.com/data-science-vs-machine-learning/). The in-memory computing technique is what makes it different from other tools and components and supports a wide variety of applications. It was developed by the Apache Software foundation in Java language primarily.
4. **R language**: [R is a programming language](https://www.educba.com/r-programming-language/)and a free software environment which is used for statistical computing and also for graphics in one of the most important languages in R. This is one among the most popular language among data scientists, data miners and data practitioners for developing statistical software and majorly in data analytics

**Let us now discuss the technologies related to Data Visualization.**

1. **Tableau:** It is the fastest and powerful growing [data visualization tool](https://www.educba.com/data-visualization-tools/) that is used in the business intelligence domain. Data analysis is a very fast machine that is possible with the help of Tableau and visualizations are created in the form of Worksheets and dashboards. It is developed by the tableau company in the year 2013 and is written in Python, C++, Java and C. Companies which are making use of Tableau are: QlikQ, Oracle Hyperion, Cognos, etc.
2. **Plotly**: Plotly is mainly used for making Graphs and associated components faster and more efficient. It has a more rich set of libraries and APIs such as MATLAB, Python, R, Arduino, Julia, etc. This can be [used interactively in Jupyter notebook](https://www.educba.com/install-jupyter-notebook/) and Pycharm and can be used to style interactive Graphs. It was first developed in 2012 and written in javascript. The few companies which are using Plotly are paladins, bitbank, etc.

Hadoop’s Parallel World

Here I am today talking about Elephant:Hadoop. We are in a world today where is there is exabytes of data being generated every data. Consider the following statistics

**Every minute:**

* Facebook users share nearly 2.5 million pieces of content.
* Twitter users tweet nearly 400,000 times.
* Instagram users post nearly 220,000 new photos.
* YouTube users upload 72 hours of new video content.
* Apple users download nearly 50,000 apps.
* Email users send over 200 million messages.
* Amazon generates over $80,000 in online sales.

Isn’t is just too vast. And thus to handle this amount of data there must be some technologies in place. In order to cope, Google invented a new style of data processing known as MapReduce. A year after Google published a white paper describing the MapReduce framework, Doug Cutting and Mike Cafarella, inspired by the white paper, created Hadoop to apply these concepts to an open-source software framework to support distribution for the Nutch search engine project. Apache Hadoop is one technology that has been the darling of Big Data talk. Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to derive the complete value from all their data.

To understand Hadoop, we must understand two fundamental things about it. They are: How Hadoop stores files, and how it processes data.Imagine we have a file that was larger than our PC’s capacity. We could not store that file, right? Hadoop lets us store files bigger than what can be stored on one particular node or server. So that we can store very, very large files. It also lets us store many, many files.

The two critical components of Hadoop are:

1. **The Hadoop Distributed File System(HDFS)**. HDFS is the storage system for a Hadoop cluster.When data lands in the cluster HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. HDFS breaks it into pieces and distributes those pieces among servers participating
2. **MapReduce.**Because Hadoop stores the entire dataset in small pieces across a collection of servers,analytical jobs can be distributed,in parallel,to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into comprehensive answer. MapReduce is the agent that distributes the work and collects the results. Map and Reduce are two functions with shuffle in between which is handled by the system.

Both HDFS and MapReduce are designed to continue to work in the face of system failure. HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails or data is damaged whether due to hardware or software problems,HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster. Likewise,when an analysis job is running,MapReduce monitors progress of each of the servers participating in the job.If one of them fails before completing its work,MapReduce automatically starts another instance of the task on another server that has copy of the data. Thus Hadoop provides scalable,reliable and fault-tolerant services for data storage and analysis at very low cost.

**Data Discovery**

Data discovery is the process of navigating or applying advanced analytics to data to detect informative patterns that could not have been discovered otherwise. Like a golfer stepping back from the ball to assess the terrain before a putt, data discovery lets businesses take a step back from individual data points, combine data from multiple sources — including external third-party data — and see the big picture, which in turn leads to better decision-making and business strategy. So, when performing data discovery, you may not always know exactly what you’re looking for — you may simply be seeking patterns and outliers to better understand your data.

Crucially, data discovery does not require business users to build elaborate models. Most companies that use data discovery do so as part of their business intelligence (BI) software, which provides them with a complete view of their organizations in a simple dashboard or visual format.

Data discovery is a five-step process. It is also an iterative process, which means companies can continue to collect, analyze and refine their data discovery approach over time by drawing on their results and feedback from business stakeholders.

* **Step 1: Identify needs.** Effective data discovery begins with a clear purpose, such as the resolution of a pain point. This means considering what kinds of data would be helpful to know, while remaining open to the unexpected insight along the way. For instance, a distributor of fast-moving consumer goods (FMCG) might decide to re-examine its logistics data in an effort to reduce food waste during shipment by 10%. Or a retail bank might analyze its web data with the aim of reducing bounce rates for new prospects.
* **Step 2: Combine data from relevant sources.** For data discovery to be effective, it is important to combine and integrate data from multiple sources because no single data stream tells the complete story. This process is sometimes referred to as data crunching.
* **Step 3: Cleanse and prep the data.** This is the heavy lifting part of data discovery — and a key part of its value. Cleaning the data and preparing it for analysis helps organizations reduce the “noise” in their data and get clearer direction from their data analyses.
* **Step 4: Analyze the data.** With information combined from multiple departments, integrated with external data and cleansed for analysis, business leaders can gain a complete view of their operations and solve the operational riddles that stand in the way of efficiency.
* **Step 5: Record learnings and iterate.** Data discovery is not a one-off process; it is a commitment to continuous improvement. In the bestseller book Outliers, author Malcolm Gladwell said it takes people 10,000 hours of practice to master a particular skill — and the same is true of businesses learning to master their data. They must treat data discovery as a way of life with the aim of improving and running more efficiently over time.

## 3 Categories of Data Discovery

Data discovery comes in multiple forms, combining analyses, modeling and visual outputs. To gain the most value from the process, businesses need to understand the interplay among their various data streams. With the help of visual discovery tools and business intelligence (BI) software, the following three categories of data discovery can help a company gain a big-picture view of its data in a single, easy-to-digest format.

### **Data Preparation**

Data preparation is a crucial step that should come before any meaningful data discovery and analysis. It involves the cleaning, reformatting and merging of data from all sources so it can be analyzed in a consistent format. In the same way a hockey player skates faster on sharpened blades, data discovery becomes more effective when businesses properly prepare their data. This might include deduplication, deleting null values, detecting outliers and any other ways to ensure only high-quality data goes through to data analysis.

### **Visual Analysis**

Visualizing data is one of the most effective ways to fully comprehend the insight it contains. Whether in the form of a [chart](https://www.netsuite.com/portal/resource/articles/erp/chart.shtml), data flow diagram or dashboard, data visualization helps those not trained in data science to understand the relationships among their various data streams in a way that feels intuitive. For example, design teams can easily learn how customers are using their products and adapt their work accordingly. And finance teams can get a snapshot of cost versus revenue for every department in the business and pinpoint areas for improvement.

### **Guided Advanced Analytics**

Guided advanced analytics combines both descriptions and visuals to paint a complete picture of a company’s data. Where typical analytics output focuses on narrow descriptions of the data itself, guided analytics allow businesses to see the wider implications of their data discovery efforts, including the relationship among data streams from different teams and processes. Guided advanced analytics is particularly valuable for businesses navigating the shift to ecommerce, where the integration of web data with existing data streams is crucial to strategic decision-making.

**Cloud and Big Data**

Big data and cloud computing are two distinctly different ideas, but the two concepts have become so interwoven that they are almost inseparable. It's important to define the two ideas and see how they relate.

Big data refers to vast amounts of data that can be structured, semistructured or unstructured. It is all about analytics and is usually derived from different sources, such as user input, IoT sensors and sales data.

Big data also refers to the act of processing enormous volumes of data to address some query, as well as identify a trend or pattern. Data is analyzed through a set of mathematical algorithms, which vary depending on what the data means, how many sources are involved and the business's intent behind the analysis. Distributed computing software platforms, such as Apache Hadoop, Databricks and Cloudera, are used to split up and organize such complex analytics.

Cloud computing provides computing resources and services on demand. A user can easily assemble the desired infrastructure of cloud-based compute instances and storage resources, connect cloud services, upload data sets and perform analyses in the cloud. Users can engage almost limitless resources across the public cloud, use those resources for as long as needed and then dismiss the environment -- [paying only for the resources](https://www.techtarget.com/searchcloudcomputing/feature/Lower-your-cloud-bill-with-these-AWS-cost-control-tips) and services that were actually used.

The public cloud has emerged as an ideal platform for big data. A cloud has the resources and services that a business can use on demand, and the business doesn't have to build, own or maintain the infrastructure. Thus, the cloud makes big data technologies accessible and affordable to almost any size of enterprise.

### **The pros of big data in the cloud**

The cloud brings a variety of important benefits to businesses of all sizes. Some of the most immediate and substantial benefits of big data in the cloud include the following.

#### **Scalability**

A typical business data center faces limits in physical space, power, cooling and the budget to purchase and deploy the sheer volume of hardware it needs to build a big data infrastructure. By comparison, a public cloud manages hundreds of thousands of servers spread across a fleet of global data centers. The infrastructure and software services are already there, and users can assemble the infrastructure for a big data project of almost any size.

#### **Agility**

Not all big data projects are the same. One project may need 100 servers, and another project might demand 2,000 servers. With cloud, users can employ as many resources as needed to accomplish a task and then release those resources when the task is complete.

#### **Cost**

A business data center is an enormous capital expense. Beyond hardware, businesses must also pay for facilities, power, ongoing maintenance and more. The cloud works all those costs into a flexible rental model where resources and services are available on demand and follow a pay-per-use model.

#### **Accessibility**

Many clouds provide a global footprint, which enables resources and services to deploy in most major global regions. This enables data and processing activity to take place proximally to the region where the big data task is located. For example, if a bulk of data is stored in a certain region of a cloud provider, it's relatively simple to implement the resources and services for a big data project in that specific cloud region -- rather than sustaining the cost of moving that data to another region.

#### **Resilience**

Data is the real value of big data projects, and the benefit of cloud resilience is in data storage reliability. Clouds replicate data as a matter of standard practice to maintain high availability in storage resources, and even more durable storage options are available in the cloud.

### **The cons of big data in the cloud**

Public clouds and many third-party big data services have proven their value in big data use cases. Despite the benefits, businesses must also consider some of the potential pitfalls. Some major disadvantages of big data in the cloud can include the following.

#### **Network dependence**

Cloud use depends on complete network connectivity from the LAN, across the internet, to the cloud provider's network. Outages along that network path can result in increased latency at best or complete cloud inaccessibility at worst. While an outage might not impact a big data project in the same ways that it would affect a mission-critical workload, the effect of outages should still be considered in any big data use of the cloud.

#### **Storage costs**

Data storage in the cloud can present a substantial long-term cost for big data projects. The three principal issues are data storage, data migration and data retention. It takes time to load large amounts of data into the cloud, and then those storage instances incur a monthly fee. If the data is moved again, there may be additional fees. Also, big data sets are often time-sensitive, meaning that some data may have no value to a big data analysis even hours into the future. Retaining unnecessary data costs money, so businesses must [employ comprehensive data retention](https://www.techtarget.com/searchcloudcomputing/tip/Its-time-to-implement-a-cloud-data-retention-policy) and deletion policies to manage cloud storage costs around big data.

#### **Security**

The data involved in big data projects can involve proprietary or personally identifiable data that is subject to data protection and other industry- or government-driven regulations. Cloud users must take the steps needed to maintain security in cloud storage and computing through adequate authentication and authorization, encryption for data at rest and in flight, and copious logging of how they access and use data.

#### **Lack of standardization**

There is no single way to architect, implement or operate a big data deployment in the cloud. This can lead to poor performance and expose the business to possible security risks. Business users should document big data architecture along with any policies and procedures related to its use. That documentation can become a foundation for optimizations and improvements for the future.

**Predictive Analytics**

The term predictive analytics refers to the use of [statistics](https://www.investopedia.com/terms/s/statistics.asp) and modeling techniques to make predictions about future outcomes and performance. Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events. Predictive analysis can also be used to improve [operational efficiencies](https://www.investopedia.com/terms/o/operationalefficiency.asp)and reduce [risk](https://www.investopedia.com/terms/r/risk.asp).

Predictive analytics is a form of technology that makes predictions about certain unknowns in the future. It draws on a series of techniques to make these determinations, including [artificial intelligence (AI)](https://www.investopedia.com/alternative-investments-4427781), [data mining](https://www.investopedia.com/terms/d/datamining.asp), machine learning, modeling, and statistics.3 For instance, data mining involves the analysis of large sets of data to detect patterns from it. Text analysis does the same, except for large blocks of text.

[Predictive models](https://www.investopedia.com/terms/p/predictive-modeling.asp) are used for all kinds of applications, including:

* Weather forecasts
* Creating video games
* Translating voice to text for mobile phone messaging
* Customer service
* Investment portfolio development

## Uses of Predictive Analytics

Predictive analytics is a decision-making tool in a variety of industries.

### **Forecasting**

Forecasting is essential in manufacturing because it ensures the optimal utilization of resources in a [supply chain](https://www.investopedia.com/terms/s/supplychain.asp). Critical spokes of the supply chain wheel, whether it is inventory management or the shop floor, require accurate forecasts for functioning.

Predictive modeling is often used to clean and optimize the quality of data used for such forecasts. Modeling ensures that more data can be ingested by the system, including from customer-facing operations, to ensure a more accurate forecast.

### **Credit**

[Credit scoring](https://www.investopedia.com/terms/c/credit_scoring.asp) makes extensive use of predictive analytics. When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to perform on any credit extended.

### **Underwriting**

Data and predictive analytics play an important role in underwriting. Insurance companies examine policy applicants to determine the likelihood of having to pay out for a future [claim](https://www.investopedia.com/terms/i/insurance_claim.asp) based on the current risk pool of similar policyholders, as well as past events that have resulted in payouts. Predictive models that consider characteristics in comparison to data about past policyholders and claims are routinely used by [actuaries](https://www.investopedia.com/terms/a/actuarial-science.asp).

### **Marketing**

Individuals who work in this field look at how consumers have reacted to the overall economy when planning on a new campaign. They can use these shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Active traders, meanwhile, look at a variety of metrics based on past events when deciding whether to buy or sell a security. Moving averages, bands, and [breakpoints](https://www.investopedia.com/terms/b/breakpoint.asp) are based on historical data and are used to forecast future price movements.

**Predictive Analysis**

To master analytics, enterprises will move from being in reactive positions (business intelligence) to forward leaning positions (predictive analytics). BIG DATA TECHNOLOGY 71 Using all the data available—traditional internal data sources combined with new rich external data sources—will make the predictions more accurate and meaningful. Because the analytics are contextual, enterprises can build confidence in the analytics and the trust will result in using analytic insights to trigger business events. By automatically triggering events, the friction in business will be greatly reduced. Algorithmic trading and supply chain optimization are just two typical examples where predictive analytics have greatly reduced the friction in business. Look for predictive analytics to proliferate in every facet of our lives, both personal and business. Here are some leading trends that are making their way to the forefront of businesses today:

* Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behaviour to recommend new purchases.
* Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.
* Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.
* Customer insight engines that integrate a wide variety of customer related info, including sentiment, behaviour, and even emotions. Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.
* Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs— or what potential competitive strategies should be used in a global business that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

**Mobile Business Intelligence**

Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one. Mobile devices are a great levelling field where making complicated actions easy is the name of the game. For example, a young child can use an iPad but not a laptop. As a result, this will drive broad-based adoption as much for the ease of use as for the mobility these devices offer. This will have an immense impact on the business intelligence sector. We interviewed Dan Kerzner, SVP Mobile at MicroStrategy, a leading provider of business intelligence software. He has been in the BI space for quite a while. People have been talking about mobile BI for quite some time, especially since the 1999 release of the good-old BlackBerry. However, it seems as though we have finally hit an inflection point. Kerzner explains his view on this topic: We have been working on Mobile BI for a while but the iPad was the inflection point where I think it started to become mainstream. I have seen customers over the past decade who focused on the mobile space generally and mobile applications in particular. One client in particular told me that he felt like he was pushing a boulder up a hill until he introduced mobility to enhance productivity. Once the new smart phones and tablets arrived, his phone was ringing off the hook and he was trying to figure out which project to say yes to, because he couldn’t say yes to everyone who suddenly wanted mobile analytics in the enterprise. That experience of folks who have been trying to use mobility for a long time to drive productivity and having really only pockets of success and then suddenly flipping over and becoming very pervasive is starting to be well understood now. In terms of why that’s the case, Dan ’s perspective on that is that with the advent of touch-driven devices, you get a set of phones that are really much more about software than they are about being a phone: You turn off the iPhone and it’s kind of a brick, nothing to it. It doesn’t look like a phone. But you turn it on and the animating experience of it is the screen and the software that flows through that screen and the usability you get from having that touch-driven device. What’s happened is suddenly you get a world where you actually have a form factor which lends itself to the power and flexibility, creativity, and innovation that comes with software development. That hadn’t really been the case before. You sort of had it with some of the Palm organizer devices that were out there and you started to have it in a light-touch way with the early Blackberries. But it was always still your phone first, your messaging, you weren’t fundamentally software driven. I think the combination of multi-touch and having a software oriented device is what has unlocked the potential of these devices to really bring mobile analytics and intelligence to a much wider audience in a productive way.

**Crowdsourcing Analytics**

In October 2006, Netflix, an online DVD rental business, announced a contest to create a new predictive model for recommending movies based on past user ratings. The grand prize was $1,000,000! While this may seem like a PR gimmick, it wasn’t. Netflix already had an algorithm to solve the problem but thought there was an opportunity to realize additional model “lift,” which would translate to huge top-line revenue. Netflix was an innovator in a space now being termed crowdsourcing. Crowdsourcing is a recognition that you can ’t possibly always have the best and brightest internal people to solve all your big problems. By creating an open, competitive environment with clear rules and goals, Netflix realized their objective and, yes, they did create a lot of buzz about their organization in the process. Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models. Let ’s face it, you can ’t “grow” a Ph.D. (or big brain) overnight. It takes years of learning and experience to get the knowledge to create algorithms and predictive models. So crowd sourcing is a way to capitalize on the limited resources that are available in the marketplace. It’s often been said that competition brings out the best in us. We are all attracted to contests; our passion for competing seems hardwired into our souls. Apparently, even predictive modelers find the siren song of competition irresistible. That’s what a small Australian fi rm, Kaggle, has discovered—when given the chance, data scientists love to duke it out, just like everyone else. Kaggle describes itself as “an innovative solution for statistical/analytics outsourcing.” That’s a very formal way of saying that Kaggle manages competitions among the world ’s best data scientists. Here ’s how it works: Corporations, governments, and research laboratories are confronted with complex statistical challenges. They describe the problems to Kaggle and provide data sets. Kaggle converts the problems and the data into contests that are posted on its web site. The contests feature cash prizes ranging in value from $100 to $3 million. Kaggle ’s clients range in size from tiny start-ups to multinational corporations such as Ford Motor Company and government agencies such as NASA. According to Anthony Goldbloom, Kaggle ’s founder and CEO, “The idea is that someone comes to us with a problem, we put it up on our website, and then people from all over the world can compete to see who can produce the best solution.” In essence, Kaggle has developed a remarkably effective global platform for crowdsourcing thorny analytic problems. What ’s especially attractive about Kaggle ’s approach is that it is truly a win-win scenario—contestants get access to real-world data (that has been carefully “anonymized” to eliminate privacy concerns) and prize sponsors reap the benefits of the contestants’ creativity. Crowdsourcing is a disruptive business model whose roots are in technology but is extending beyond technology to other areas. There are various types of crowdsourcing, such as crowd voting, crowd purchasing, wisdom of crowds, crowd funding, and contests. Take for example: ■ 99designs.com/ , which does crowdsourcing of graphic design ■ agentanything.com/ , which posts “missions” where agents vie for to run errands ■ 33needs.com/ , which allows people to contribute to charitable programs that make a social impact

**Inter – and Trans-Firewall Analytics**

Over the last 100 years, supply chains have evolved to connect multiple companies and enable them to collaborate to create enormous value to the end consumer via concepts such as CPFR, VMI, and so on. Decision science is witnessing a similar trend as enterprises are beginning to collaborate on insights across the value chain. For instance, in the health care industry, rich consumer insights can be generated by collaborating on data and insights from the health insurance provider, pharmacy delivering the drugs, and the drug manufacturer. In-fact, this is not necessarily limited to companies within the traditional demand-supply value chain. For example, there are instances where a retailer and a social media company can come together to share insights on consumer behaviour that will benefit both players. Some of the more progressive companies are taking this a step further and working on leveraging the large volumes of data outside the firewall such as social data, location data, and so forth. In other words, it will be not very long before internal data and insights from within the firewall is no longer a differentiator. We see this trend as the move from intra- to inter- and trans-firewall analytics. Yesterday companies were doing functional silo-based analytics. Today they are doing intra-firewall analytics with data within the firewall. Tomorrow they will be collaborating on insights with other companies to do inter-firewall analytics as well as leveraging the public domain spaces to do trans-firewall analytics. Setting up inter-firewall and trans-firewall analytics can add significant value. However it does present some challenges. First, as one moves outside the firewall, the information-to-noise ratio increases, putting additional requirements on analytical methods and technology requirements. Further, organizations are often limited by a fear of collaboration and an overreliance on proprietary information. The fear of collaboration is mostly driven by competitive fears, data privacy concerns, and proprietary orientations that limit opportunities for cross-organizational learning and innovation. While it is clear that the transition to an inter- and trans-firewall paradigm is not easy, we feel it will continue to grow and at some point it will become a key weapon, available for decisions scientists to drive disruptive value and efficiencies.