Unit 1

Understa… and its
importan… analytics and its
classificat… medicine –
advertisi…

1.1.1 An Introduction to Big Data Analytics

Big data analytics can be defined as a process of collecting, organizing and analyzing large and varied data sets using advanced analytics techniques to uncover the hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. This helps the organizations to make informed decisions.

To understand Big Data Analytics you have to first understand What Analytics is?

Analytics : Analytics is an encompassing and multidimensional field. It uses mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data.

Big Data Analytics: Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it immediately. Big Data Analytics helps you to understand your organization better. With the use of Big data analytics, one can make the informed decisions without blindly relying on guesses.

The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence(BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things.

Examples Of Big Data

Big data works on the data produced by various devices and their applications.

1) **Black Box Data:** It is an incorporated by flight crafts, which stores a large sum of information, which includes the conversation between crew members and any other communications (alert messages or any order passed)by the technical grounds duty staff.

A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

2) **Social Media Data**: Social networking sites such as Face book and Twitter contains the information and the views posted by millions of people across the globe.

   The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

3) **Stock Exchange Data**: It holds information (complete details of in and out of business transactions) about the 'buyer' and 'seller' decisions in terms of share between different companies made by the customers.

   The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.

4) **Power Grid Data**: The power grid data mainly holds the information consumed by a particular node in terms of base station.

5) **Transport Data**: It includes the data's from various transport sectors such as model, capacity, distance and availability of a vehicle.

6) **Search Engine Data**: Search engines retrieve a large amount of data from different sources of database.

Big data facts:

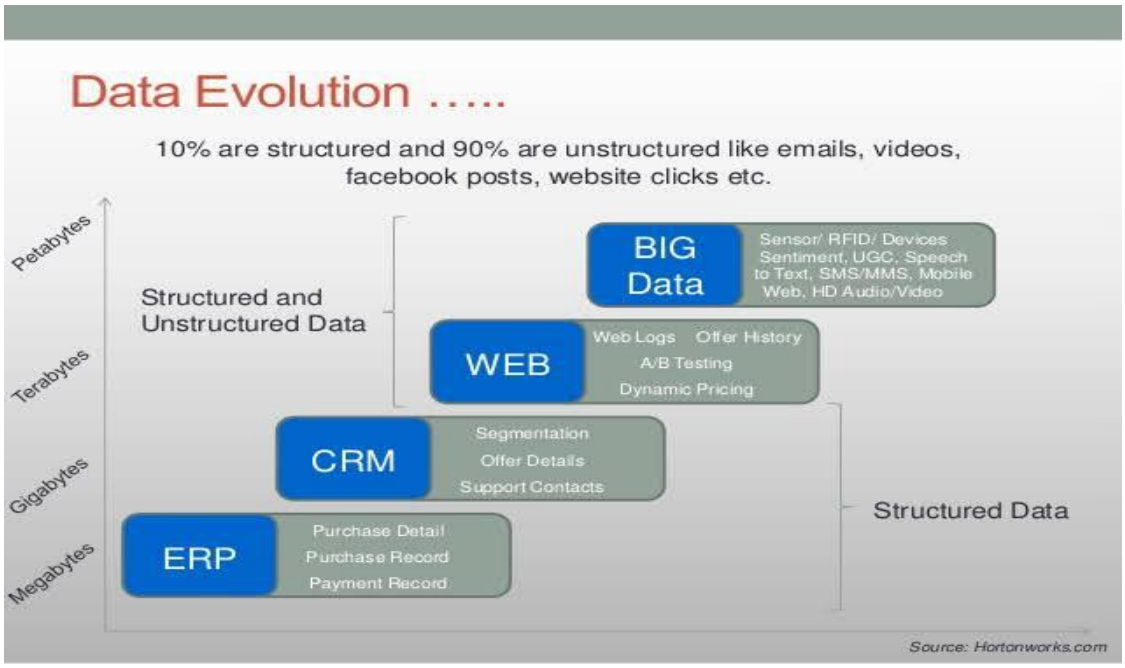
Data is everywhere. According to the predictions, by 2025, 463 exabytes of data will be generated each day which is equal to 212,765,957 DVDs per day!
Billions of emails, millions of tweets are sent every day. Thus, every person is generating a huge amount of data which is rising exponentially.
The Big Data market will sooner reach $103B by 2023

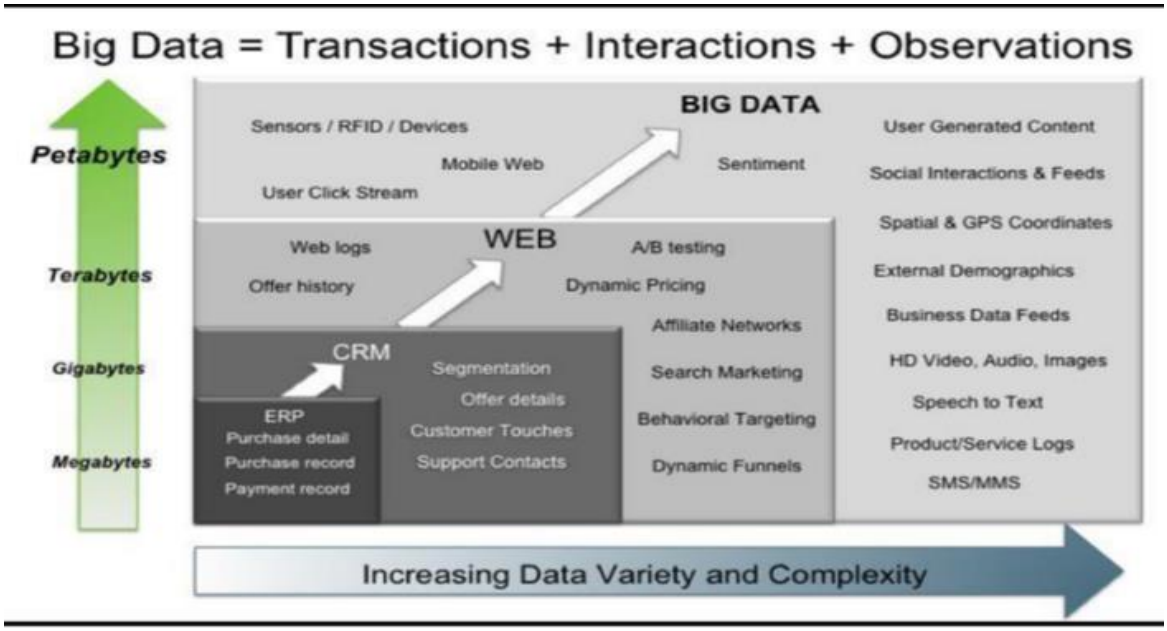
- We create 2.5 quintillion ($1x10^{18}$) bytes every day
- 90% of world's data was created in the last 2 years
- 80% of world's data is unstructured
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data
- Facebook processes 500TB per day
- 72 hours of video are uploaded to youtube every minute
- Over 5 billion people use cell phones to call, send SMS, email, browse Internet, and interact via social networking sites.

1.1.2 History and Evolution of Big Data Analytics:

The concept of big data has been around for years; most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it. But even in the 1950s, decades before anyone uttered the term "big data," businesses were using basic analytics essentially numbers in a spreadsheet that were manually examined to uncover insights and trends.

The new benefits that big data analytics brings to the table, however, are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and unearthed information that could be used for future decisions, today that business can identify insights for immediate decisions. The ability to work faster – and stay agile – gives organizations a competitive edge they didn't have before.





1.1.3 Values of Big Data Analytics:

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. Here are the most important values of Big Data,

1. **Cost reduction:** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making:** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
3. **New products and services:** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

1.1.4 Advantages of Big Data (Features)

- One of the biggest advantages of Big Data is predictive analysis. Big Data analytics tools can predict outcomes accurately, thereby, allowing businesses and organizations to make better decisions, while simultaneously optimizing their operational efficiencies and reducing risks.
- By harnessing data from social media platforms using Big Data analytics tools, businesses around the world are streamlining their digital marketing strategies to enhance the overall consumer experience. Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.
- Being accurate, Big Data combines relevant data from multiple sources to produce highly actionable insights. Almost 43% of companies lack the necessary tools to

filter out irrelevant data, which eventually costs them millions of dollars to hash out useful data from the bulk. Big Data tools can help reduce this, saving you both time and money.

- Big Data analytics could help companies generate more sales leads which would naturally mean a boost in revenue. Businesses are using Big Data analytics tools to understand how well their products/services are doing in the market and how the customers are responding to them. Thus, the can understand better where to invest their time and money.
- With Big Data insights, you can always stay a step ahead of your competitors. You can screen the market to know what kind of promotions and offers your rivals are providing, and then you can come up with better offers for your customers. Also, Big Data insights allow you to learn customer behavior to understand the customer trends and provide a highly 'personalized' experience to them.

1.1.5 Applications of Big Data analytics across different industries:

1. **Banking**

The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.

Large amounts of information will be streaming in into banks, managing all this data and getting proper insights would be possible only with big data analytics. This is important to understand customers and boost their satisfaction, and also to minimize risk and fraud.

2. **Government**

When government agencies are able to harness and apply analytics to their big data, they gain significant ground when it comes to managing utilities, running agencies, dealing with traffic congestion or preventing crime.

3. **Health Care**

Big Data has already started to create a huge difference in the healthcare sector. With the help of predictive analytics, medical professionals and HCPs are now able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring – all powered by Big Data and AI – are helping change lives for the better.

Patient records, Treatment plans, Prescription information. When it comes to health care, everything needs to be done quickly, accurately. And, in some cases, with enough transparency to satisfy stringent industry regulations. When big data is managed effectively, health care providers can uncover hidden insights that improve patient care.

4. **Education**

Big Data is also helping enhance education today. Education is no more limited to the physical bounds of the classroom – there are numerous online educational courses to learn

from. Academic institutions are investing in digital courses powered by Big Data technologies to aid the all-round development of budding learners.

Educators armed with data-driven insight can make a significant impact on school systems, students, and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress, and can implement a better system for evaluation and support of teachers and principals.

5. **Manufacturing**

According to TCS Global Trend Study, the most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality. In the manufacturing sector, Big data helps create a transparent infrastructure, thereby, predicting uncertainties and incompetencies that can affect the business adversely.

Armed with insight that big data can provide, manufacturers can boost quality and output while minimizing waste – processes that are key in today's highly competitive market. More and more manufacturers are working in an analytics-based culture, which means they can solve problems faster and make more agile business decisions.

6. **Retail**

Customer relationship building is critical to the retail industry. And the best way to manage that is to manage big data. Retailers need to know the best way to market to customers. The most effective way to handle transactions, and the most strategic way to bring back lapsed business. Big data remains at the heart of all those things.

7. **IT**

One of the largest users of Big Data, IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity, and minimize risks in business operations. By combining Big Data technologies with ML and AI, the IT sector is continually powering innovation to find solutions even for the most complex of problems.

1.2 Characteristics of Big Data

Gartner analyst Doug Laney listed the **3 'V's of Big Data – Variety, Velocity, and Volume.**

According to Gartner, the definition of Big Data –

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

1) Variety

Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in

the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

Types of Big Data

a. **Structured**

By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc.,** will be present in an organized manner.

However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes. (***10^{21} bytes* = *1 zettabyte*** = ***one billion terabytes***) Looking at these figures one can easily understand why the name Big Data is given and imagine the challenges involved in its storage and processing.

Examples Of Structured Data :

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

b. **Unstructured**

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data.

A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples Of Un-structured Data

The output returned by 'Google Search' , Email.

c. **Semi-structured**

Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

Example of semi-structured data is a data represented in an XML file.

Personal data stored in an XML file-

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

Data Sources

a. Enterprise data
    - Serves business objectives, well defined
    - Customer information
    - Transactions, e.g. Purchases

b. Experimental/Observational data (EOD)
    - Created by machines from sensors/devices
    - Trading systems, satellites
    - Microscopes, video streams, Smart meters

1. Social media
    - Created by humans
    - Messages, posts, blogs, Wikis

2) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

Examples

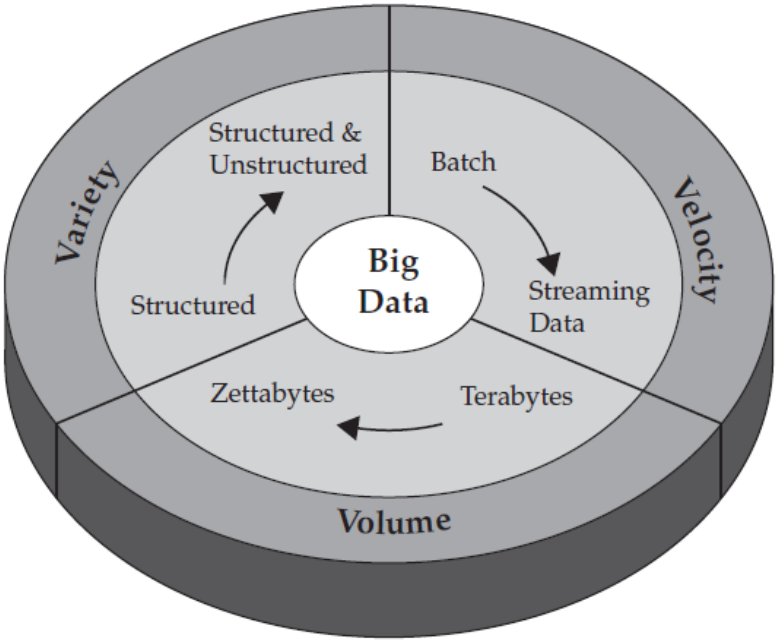
- LHC (Cern) with all experiments about 25 GB/s
  (The *Large Hadron Collider* (*LHC*) is the world's largest and most powerful particle accelerator.)
- The **Square Kilometre Array** (SKA- is a radio telescope) Array 700 TB/s (in 2018)
- 50k Google searches per sec
- Facebook 30 Billion content pieces shared per month
- Youtube 72 hours of video per min.

3) Volume

We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses.

Examples

- Wikipedia corpus with history ca. 10 TByte
- Wikimedia commons ca. 23 TByte
- Google search index ca. 46 Gigawebpages
- YouTube per year 76 PByte (2012)



(iv) Variability/ **Veracity –** Trustworthiness of Data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively. *Veracity* referrers to the data quality, relevance, uncertainty, reliability and predictive value, while *variability* regards about consistency of the data over time.

Examples

- Data involves some uncertainty and ambiguities
- Mistakes can be introduced by humans and machines
- People sharing accounts
- Like it today, dislike it tomorrow
- Wrong system timestamps
- A Twitter post has hashtags, typos and abbreviations.

Data Quality is vital!

Analytics and conclusions rely on good data quality

Garbage data + perfect model => garbage results

Perfect data + garbage model => garbage results

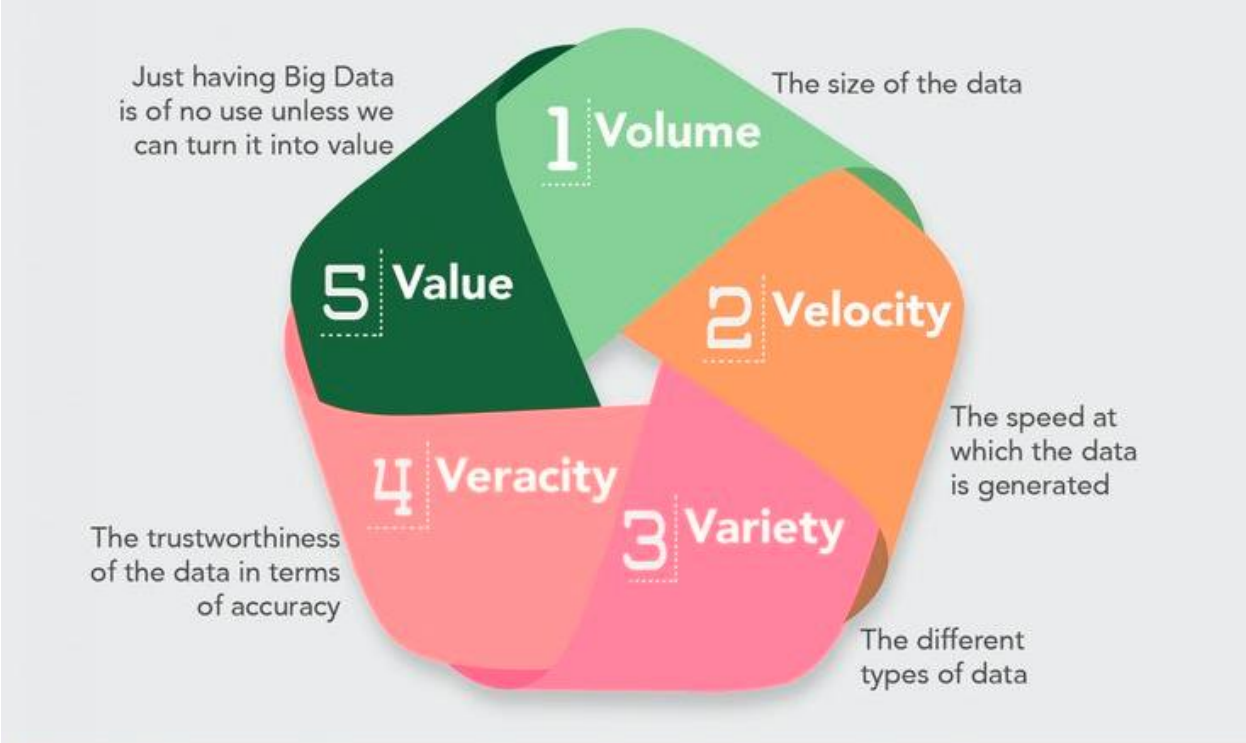
GIGO paradigm: *Garbage In – Garbage Out*

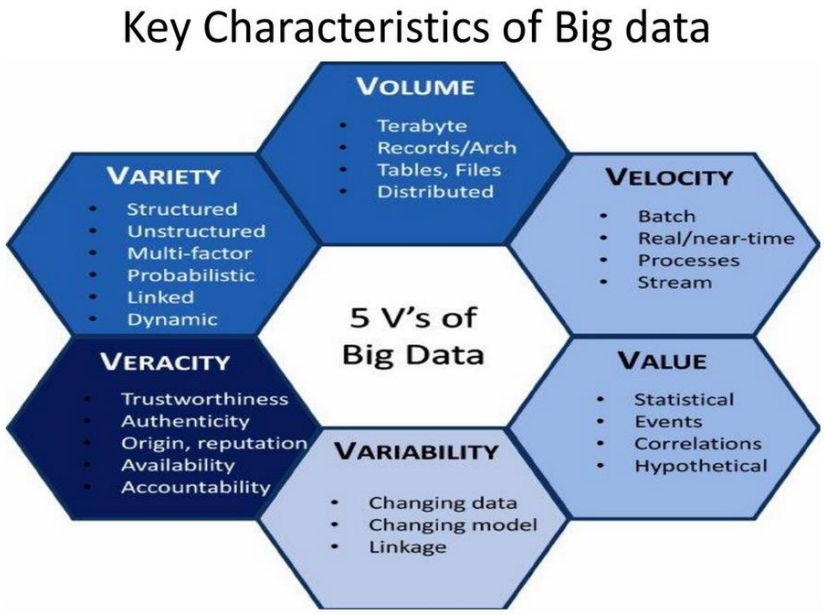
You can see that few values are missing in the below table

Department	Year	Minimum sales	Maximum sales
1	2010	?	1500
2	2011	10000	?

Data available can sometimes get messy and maybe difficult to trust. With wide variety in big data types generated, quality and accuracy are difficult to control.

(v) Value of Data: Jus having Big Data is of no use unless we can turn it into value. Raw data of Big Data is of low value. Analytics and theory about the data increases the value. Analytics transform big data into smart data.



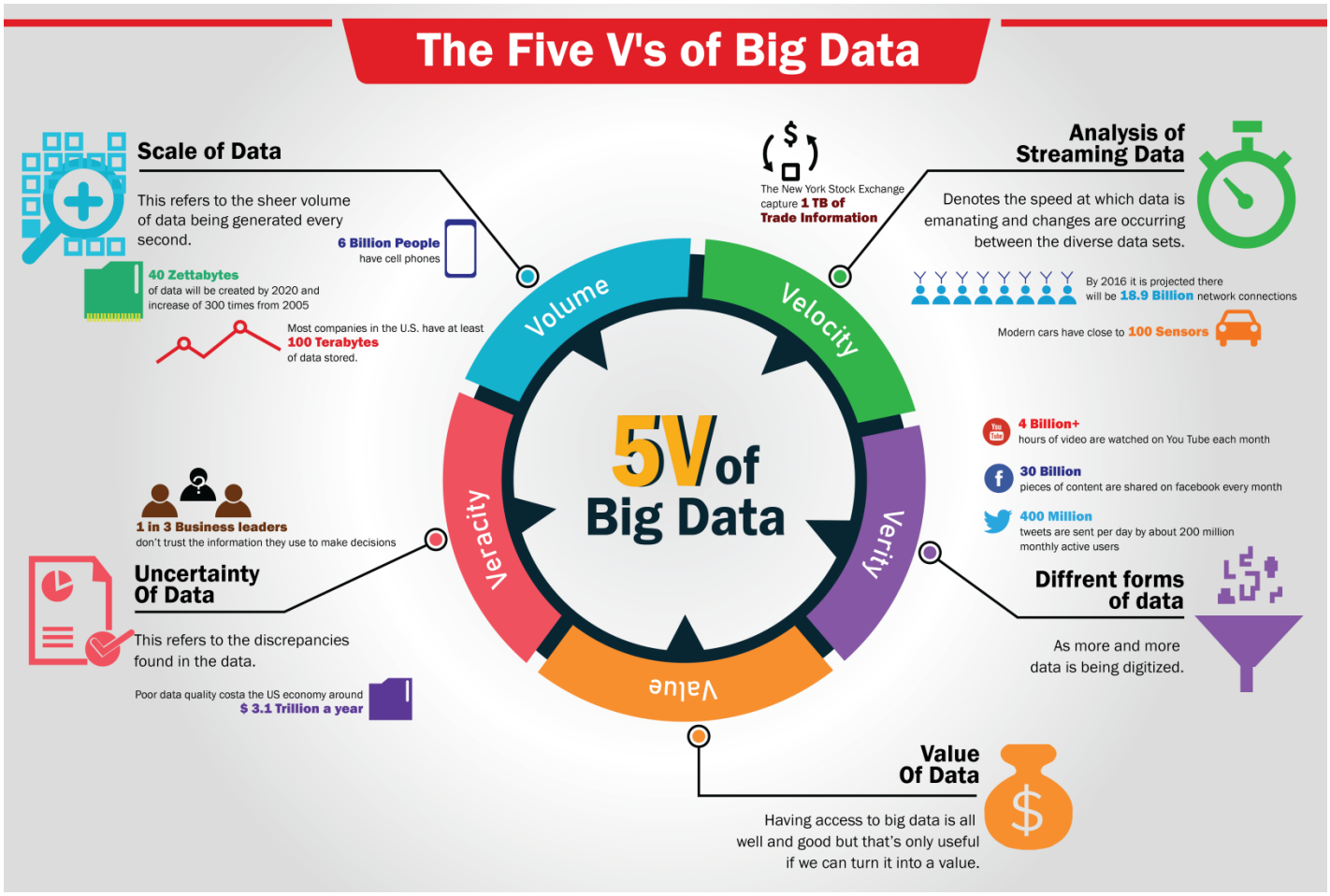


1.3. Challenges posed by Big Data:

Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.



These three characteristics cause many of the challenges that organizations encounter in their big data initiatives. Some of the most common of those big data challenges include the following:

1. Dealing with data growth / Handling a Large Amount of Data

There is a huge explosion in the data available. Look back a few years, and compare it with today, and you will see that there has been an exponential increase in the data that enterprises can access. They have data for everything, right from what a consumer likes, to how they react, to a particular scent, to the amazing restaurant that opened up in Italy last weekend.

This data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data. The most obvious challenge associated with big data is simply storing and analyzing all that information. In its Digital Universe report, IDC estimates that the amount of information stored in the world's IT systems is doubling about every two years. By 2020, the total amount will be enough to fill a stack of tablets that reaches from the earth to the moon 6.6 times. And enterprises have responsibility or liability for about 85 percent of that information.

Much of that data is unstructured, meaning that it doesn't reside in a database. Documents, photos, audio, videos and other unstructured data can be difficult to search and analyze.

Some of the newest ways developed to manage this data are a hybrid of relational databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

Netflix is a content streaming platform based on Node.js. With the increased load of content and the complex formats available on the platform, they needed a stack that could handle the storage and retrieval of the data. They used the MEAN stack, and with a relational database model, they could in fact manage the data.

It's no surprise, then, that the IDG report found, "Managing unstructured data is growing as a challenge – rising from 31 percent in 2015 to 45 percent in 2016."

In order to deal with data growth, organizations are turning to a number of different technologies. When it comes to storage, converged and hyperconverged infrastructure and software-defined storage can make it easier for companies to scale their hardware. And technologies like compression, deduplication and tiering can reduce the amount of space and the costs associated with big data storage.

On the management and analysis side, enterprises are using tools like NoSQL databases, Hadoop, Spark, big data analytics software, business intelligence applications, artificial intelligence and machine learning to help them comb through their big data stores to find the insights their companies need.

2. Generating insights in a timely manner / Real-time can be Complex

When I say data, I'm not limiting this to the "stagnant" data available at common disposal. A lot of data keeps updating every second, and organizations need to be aware of that too. For instance, if a retail company wants to analyze customer behavior, real-time data from their current purchases can help. There are Data Analysis tools available for the same – Veracity and Velocity. They come with ETL engines, visualization, computation engines, frameworks and other necessary inputs.

It is important for businesses to keep themselves updated with this data, along with the "stagnant" and always available data. This will help build better insights and enhance decision-making capabilities.

However, not all organizations are able to keep up with real-time data, as they are not updated with the evolving nature of the tools and technologies needed. Currently, there are a few reliable tools, though many still lack the necessary sophistication.

Of course, organizations don't just want to store their big data — they want to use that big data to achieve business goals. According to the NewVantage Partners survey, the most common goals associated with big data projects included the following:

1. Decreasing expenses through operational cost efficiencies

2. Establishing a data-driven culture
3. Creating new avenues for innovation and disruption
4. Accelerating the speed with which new capabilities and services are deployed
5. Launching new product and service offerings

All of those goals can help organizations become more competitive — but only if they can extract insights from their big data and then act on those insights quickly. PwC's Global Data and Analytics Survey 2016 found, "Everyone wants decision-making to be faster, especially in banking, insurance, and healthcare."

To achieve that speed, some organizations are looking to a new generation of ETL and analytics tools that dramatically reduce the time it takes to generate reports. They are investing in software with real-time analytics capabilities that allows them to respond to developments in the marketplace immediately.

3. Recruiting and retaining big data talent / Shortage of Skilled People

There is a definite shortage of skilled Big Data professionals available at this time. This has been mentioned by many enterprises seeking to better utilize Big Data and build more effective Data Analysis systems. There is a lack experienced people and certified Data Scientists or Data Analysts available at present, which makes the “number crunching” difficult, and insight building slow.

Again, training people at entry level can be expensive for a company dealing with new technologies. Many are instead working on automation solutions involving Machine Learning and Artificial Intelligence to build insights, but this also takes well-trained staff or the outsourcing of skilled developers.

In order to develop, manage and run those applications that generate insights, organizations need professionals with big data skills. That has driven up demand for big data experts — and big data salaries have increased dramatically as a result.

The 2017 Robert Half Technology Salary Guide reported that big data engineers were earning between $135,000 and $196,000 on average, while data scientist salaries ranged from $116,000 to $163, 500. Even business intelligence analysts were very well paid, making $118,000 to $138,750 per year.

In order to deal with talent shortages, organizations have a couple of options. First, many are increasing their budgets and their recruitment and retention efforts. Second, they are offering more training opportunities to their current staff members in an attempt to develop the talent they need from within. Third, many organizations are looking to technology. They are buying analytics solutions with self-service and/or machine learning capabilities. Designed to be used by professionals without a data science degree, these tools may help organizations achieve their big data goals even if they do not have a lot of big data experts on staff.

4. Integrating disparate data sources

The variety associated with big data leads to challenges in data integration. Big data comes from a lot of different places — enterprise applications, social media streams, email systems, employee-created documents, etc. Combining all that data and reconciling it so that it can be used to create reports can be incredibly difficult. Vendors offer a variety of ETL and data integration tools designed to make the process easier, but many enterprises say that they have not solved the data integration problem yet.

In response, many enterprises are turning to new technology solutions. In the IDG report, 89 percent of those surveyed said that their companies planned to invest in new big data tools in the next 12 to 18 months. When asked which kind of tools they were planning to purchase, integration technology was second on the list, behind data analytics software.

5. Validating data

Closely related to the idea of data integration is the idea of data validation. Often organizations are getting similar pieces of data from different systems, and the data in those different systems doesn't always agree. For example, the ecommerce system may show daily sales at a certain level while the enterprise resource planning (ERP) system has a slightly different number. Or a hospital's electronic health record (EHR) system may have one address for a patient, while a partner pharmacy has a different address on record.

The process of getting those records to agree, as well as making sure the records are accurate, usable and secure, is called data governance. And in the AtScale 2016 Big Data Maturity Survey, the fastest-growing area of concern cited by respondents was data governance.

Solving data governance challenges is very complex and is usually requires a combination of policy changes and technology. Organizations often set up a group of people to oversee data governance and write a set of policies and procedures. They may also invest in data management solutions designed to simplify data governance and help ensure the accuracy of big data stores — and the insights derived from them.

6. Securing big data / Data Security

A lot of organizations claim that they face trouble with Data Security. This happens to be a bigger challenge for them than many other data-related problems. The data that comes into enterprises is made available from a wide range of sources, some of which cannot be trusted to be secure and compliant within organizational standards.

They need to use a variety of data collection strategies to keep up with data needs. This in turn leads to inconsistencies in the data, and then the outcomes of the analysis. A simple example such as annual turnover for the retail industry can be different if analyzed from different sources of input. A business will need to adjust the differences, and narrow it down to an answer that is valid and interesting.

This data is made available from numerous sources, and therefore has potential security problems. You may never know which channel of data is compromised, thus compromising the security of the data available in the organization, and giving hackers a chance to move in.

It's necessary to introduce Data Security best practices for secure data collection, storage and retrieval.

Security is also a big concern for organizations with big data stores. After all, some big data stores can be attractive targets for hackers or advanced persistent threats (APTs).

However, most organizations seem to believe that their existing data security methods are sufficient for their big data needs as well. In the IDG survey, less than half of those surveyed (39 percent) said that they were using additional security measure for their big data repositories or analyses. Among those who do use additional measures, the most popular include identity and access control (59 percent), data encryption (52 percent) and data segregation (42 percent).

7. Organizational resistance

It is not only the technological aspects of big data that can be challenging — people can be an issue too.

In the NewVantage Partners survey, 85.5 percent of those surveyed said that their firms were committed to creating a data-driven culture, but only 37.1 percent said they had been successful with those efforts. When asked about the impediments to that culture shift, respondents pointed to three big obstacles within their organizations:

- Insufficient organizational alignment (4.6 percent)
- Lack of middle management adoption and understanding (41.0 percent)
- Business resistance or lack of understanding (41.0 percent)

In order for organizations to capitalize on the opportunities offered by big data, they are going to have to do some things differently. And that sort of change can be tremendously difficult for large organizations.

The PwC report recommended, "To improve decision-making capabilities at your company, you should continue to invest in strong leaders who understand data's possibilities and who will challenge the business."

One way to establish that sort of leadership is to appoint a chief data officer, a step that NewVantage Partners said 55.9 percent of Fortune 1000 companies have taken. But with or without a chief data officer, enterprises need executives, directors and managers who are going to commit to overcoming their big data challenges, if they want to remain competitive in the increasing data-driven economy.

1.1.5 Big Data Analytics and its classification

Analytics is the discovery and communication of meaningful patterns in data. Especially, valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operation research to qualify performance. Analytics often favors data visualization to communicate insight.

Firms may commonly apply analytics to business data, to describe, predict and improve business performance. Especially, areas within include predictive analytics, enterprise decision management etc. Since analytics can require extensive computation(because of big data), the algorithms and software used to analytics harness the most current methods in computer science.

In a nutshell, analytics is the scientific process of transforming data into insight for making better decisions. The goal of Data Analytics is to get actionable insights resulting in smarter decision and better business outcomes.

And it can help answer the following types of questions:

- What actually happened?
- How or why did it happen?
- What's happening now?
- What is likely to happen next?

There are four type of data analytics:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)
4. Diagnostic analytics

Predictive Analytics: Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring.
Predictive analytics holds a variety of statistical technique from modeling, machine, learning, data mining and game theory that analyze current and historical facts to make prediction about future event.

There are three basic cornerstones of predictive analytics-

- Predictive modeling
- Decision Analysis and optimization
- Transaction profiling

Predictive analytics is all about forecasting. Whether it's the likelihood of an event happening in future, forecasting a quantifiable amount or estimating a point in time at which something might happen – these are all done through predictive models.

Predictive models typically utilise a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict (e.g. the older a person, the more susceptible they are to a heart-attack – we would say that age has a linear correlation with heart-attack risk). These data are then compiled together into a score or prediction.

In a world of great uncertainty, being able to predict allows one to make better decisions. Predictive models are some of the most important utilised across a number of fields.

Descriptive Analytics: Descriptive analytics looks at data and analyze past event for insight as how to approach future events. It looks at the past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all the management reporting such as sales, marketing, operations, and finance uses this type of analysis.
Descriptive model quantifies relationship in data in a way that is often used to classify customers or prospect into groups. Unlike predictive model that focuses on predicting the behavior of single customer, Descriptive analytics identify many different relationships between customer and product.

An examples of this could be a monthly profit and loss statement. Similarly, an analyst could have data on a large population of customers. Understanding demographic information on their customers (e.g. 30% of our customers are self-employed) would be categorised as "descriptive analytics". Utilising effective visualisation tools enhances the message of descriptive analytics.

Prescriptive Analytics: Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make prediction and then suggests decision option to take advantage of the prediction.
Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefit from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when happen but also why it will happen. Further, Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.
For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography etc.

The next step up in terms of value and complexity is the prescriptive model. The prescriptive model utilises an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take. Prescriptive analysis is typically not just with one individual action, but is in fact a host of other actions.

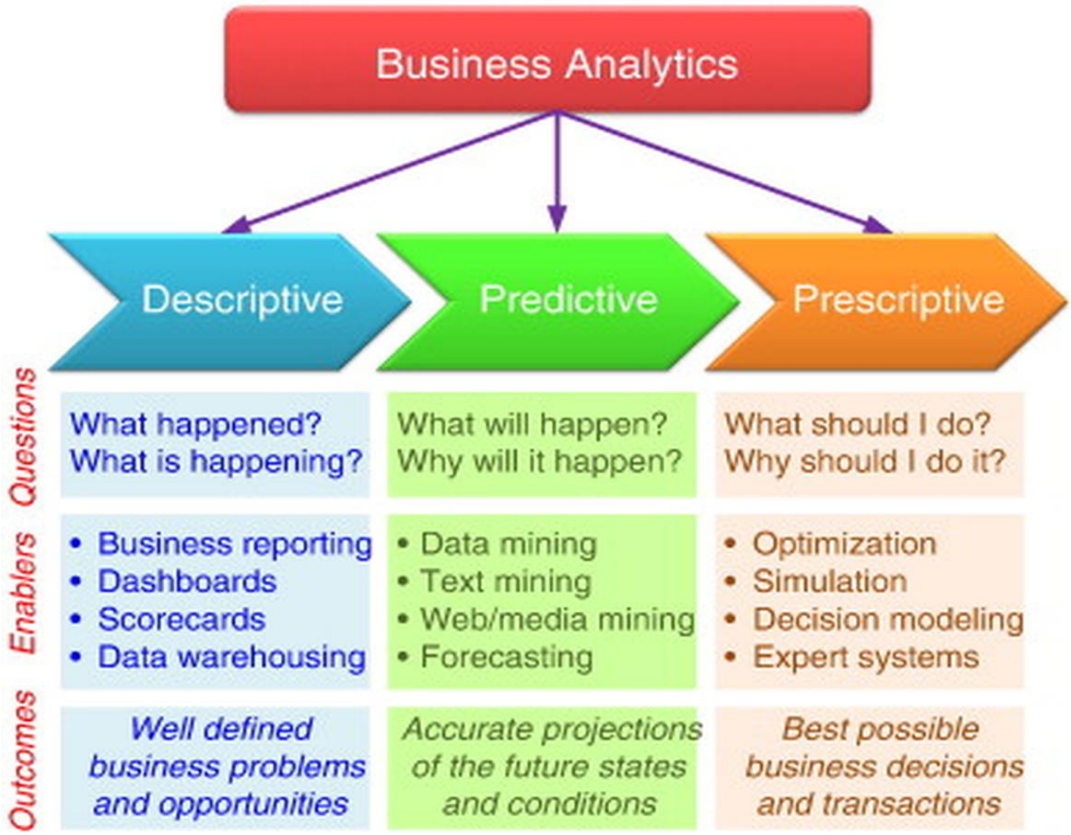
A good example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and, crucially, the current traffic constraints.

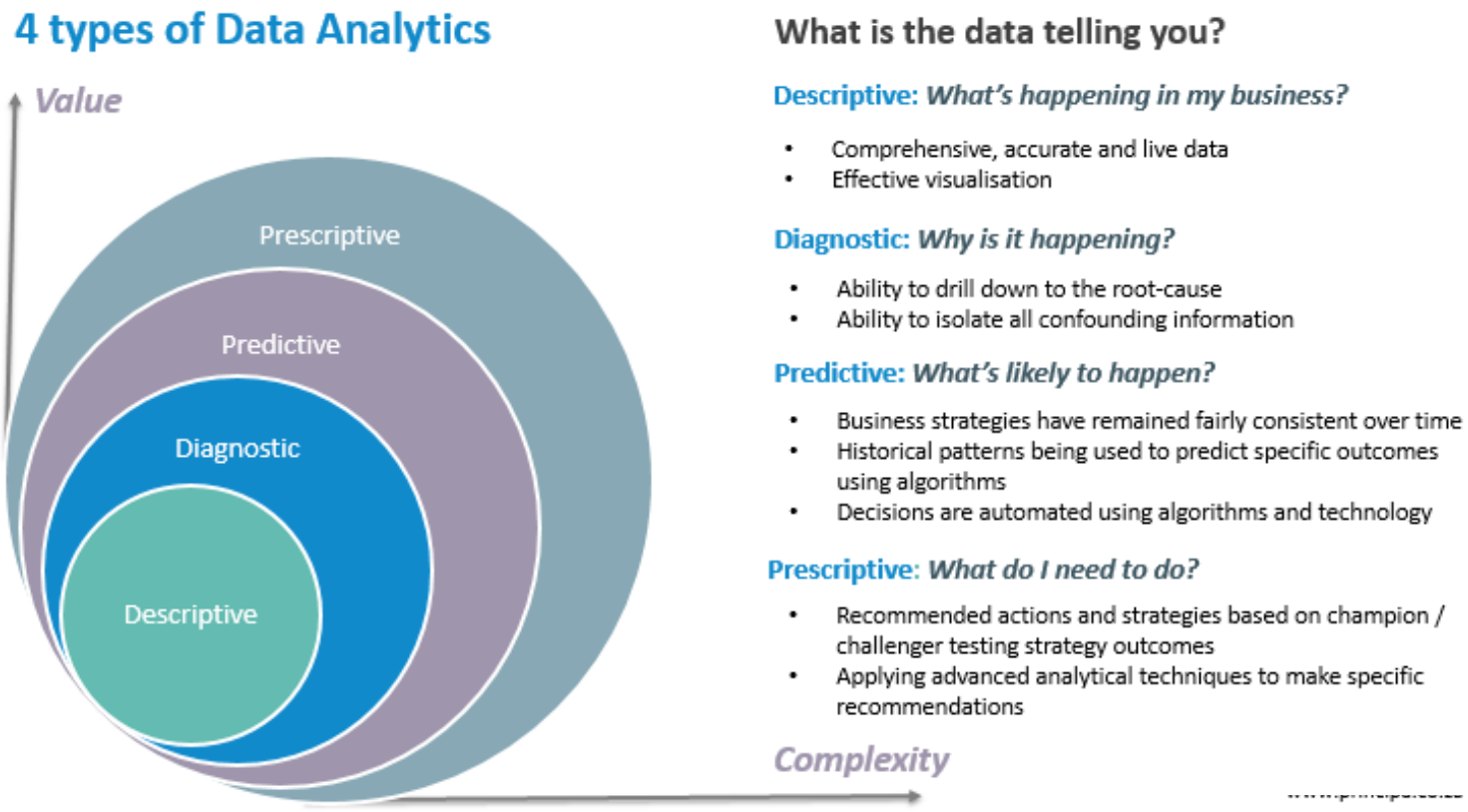
Another example might be producing an exam time-table such that no students have clashing schedules.

Diagnostic Analytics: In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.
For example, companies go for this analysis because it gives a great insight for a problem, and they also keep detailed information about there disposal otherwise data collection may turn out individual for every problem and it will be very time-consuming.

This is the next step of complexity in data analytics to descriptive analytics. On assessment of the descriptive data, diagnostic analytical tools will empower an analyst to drill down and in so doing isolate the root-cause of a problem.

Well-designed business information (BI) dashboards incorporating reading of time-series data (i.e. data over multiple successive points in time) and featuring filters and drill down capability allow for such analysis.





1.6. Big Data Applications in Healthcare

Big data is the buzzword today. It is heard everywhere, especially in the healthcare industry. Traditionally, the huge amount of data generated by the healthcare industry was stored as hard copy. This data has the capability to support a wide range of healthcare and medical functions. The digitization of such data is called Big Data. The entirety of data that is related to patient health care and well-being makes up big data. A 2011 McKinsey report estimated that the health care industry could potentially realize $300 billion in annual value by leveraging big data.

The wide diversity of big data and the pace at which it is managed makes it overwhelming. It includes clinical data from CPOE and clinical decision support systems (physicians written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine generated/sensor data, such as from monitoring vital signs; social media posts, including Twitter feeds (so-called tweets), blogs, status updates on Facebook and other platforms, and web pages; and less patient-specific information, including emergency care data, news feeds, and articles in medical journals.

Types Of Data In A Healthcare System:

Below are the types of data that gets created in a hospital or clinic:

- Clinical data: Notes made by doctors, nurses and pharmacists, prescriptions, reports from medical imaging, laboratory, pharmacy and insurance departments
- Patient data: Includes electronic patient records (EPRs)
- Machine generated/sensor data: Derived from monitoring vital signs, social media posts, website traffic, emergency care data, news feeds, and articles in medical journals.

The World Health Organization has confirmed that over the past decade maintenance of electronic health records in clinics and hospitals worldwide has literally transformed healthcare. Having structured data that can be easily tracked aids in improving all aspects of patient care as well as enhances clinical decision-making.

Sources of Big Data in Medicine

These are traditional sources of clinical data that health care providers are accustomed to viewing.

- **Electronic health records** (EHRs) collect, store, and display information such as demographics, past medical history, active medical problems, immunizations, allergies, medications, vital signs, results from laboratory and radiology tests, pathology reports, progress notes created by health care providers, and administrative and financial documents
- **Electronic medical records (EMRs)** are not identical to EHRs and usually pertain to data stored with a particular physician.
    - Health information exchanges serve as hubs between disparate clinical information systems
    - Patient registries, maintained by health care organizations on their own patients, are often linked to the EHR. Other registries track immunizations, cancer, trauma, and other public health issues on a wider geographic scale.
- **Patient portals** allow patients to access personal health information stored in a health care organization's EHR. Some patient portals also allow users to request prescription refills and exchange secure electronic messages with the health care team.

Clinical data warehouses aggregate patient-level data from multiple clinical information systems, such as EHRs and other sources listed above

Big data is very useful in the healthcare industry. Over the past decade, electronic health records (EHR) have been widely adopted in hospitals and clinics worldwide. Important clinical knowledge and a deeper understanding of patient disease patterns can be studied from such data. It will help to improve patient care and improve efficiency. There are few targeted applications of big data, such as

1. Healthcare Data solutions

With the help of big data, the vast amount of data can be stored systematically. Now doctors and other healthcare practitioners can make informed decisions as they have access to a wide range of data. Of course, the data generated will grow by leaps and bounds, and newer systems will be able to process it quickly and cost effectively.

2. Big data to fight cancer

Cancer has already become one of the leading causes of mortality and morbidity across the world today. With predictive analytics, pre-existing conditions and habit patterns can be used to foresee how vulnerable an individual is to cancer. For healthcare providers, big data has empowered them to detect and diagnose even the rarest forms of cancer at an early stage itself.

Data collected from a patient's genetic makeup have also been useful in determining treatment options that have minimal side effects. This is one among the various humanistic big data applications in healthcare.

3. **Monitoring patient vitals**

The application of big data makes it easier for hospital staff to work more efficiently. Sensors are used besides patient beds to continuously monitor blood pressure, heartbeat and respiratory rate. Any change in pattern is quickly alerted to doctors and healthcare administrators.

4. **Smoother Hospital Administration**

Healthcare administration becomes much smoother with the help of big data. It helps to reduce the cost of care measurement, provide the best clinical support, and manage the population of at-risk patients. It also helps medical experts analyze data from diverse sources. It helps healthcare providers conclude the deviations among patients and the effects treatments have on their health.

5. **Healthcare Intelligence**

Big Data can be used for healthcare Intelligence application. This will help hospitals, payers and healthcare agencies augment their competitive advantages by developing smart business solutions.

6. **Fraud Prevention and Detection**

Big data helps to prevent a wide range of errors on the side of health administrators in the form of wrong dosage, wrong medicines, and other human errors. It will also be particularly useful to insurance companies. They can prevent a wide range of fraudulent claims of insurance.

Healthcare environments are also susceptible to human error. The presence Big data helps to prevent a wide range of errors that could happen during storing, sharing and sorting different types of hospital records. Everything from faulty prescriptions and investigations to fraudulent claims of insurance can be curbed in setups that are powered by big data. on the side of health administrators in the form of wrong dosage, wrong medicines, and other human errors. It will also be particularly useful to insurance companies. They can prevent a wide range of fraudulent claims of insurance.

7. **Promotes Business Development**

For hospitals that are a part of corporate groups and medical tourism, big data can be used to frame business plans with that boost success rates. Smart business solutions aid in keeping hospitals as a profitable ecosystem for the staff and management. This will, in turn, increase the quality of care delivered to patients.

8. Telemedicine

Telemedicine has been present on the market for over 40 years, but only today, with the arrival of online video conferences, smartphones, wireless devices, and wearables, has it been able to come into full bloom. The term refers to delivery of remote clinical services using technology.

It is used for primary consultations and initial diagnosis, remote patient monitoring, and medical education for health professionals. Some more specific uses include telesurgery – doctors can perform operations with the use of robots and high-speed real-time data delivery without physically being in the same location with a patient.

Clinicians use telemedicine to provide personalized treatment plans and prevent hospitalization or re-admission. Such use of healthcare data analytics can be linked to the use of predictive analytics as seen previously. It allows clinicians to predict acute medical events in advance and prevent deterioration of patient's conditions.

By keeping patients away from hospitals, telemedicine helps to reduce costs and improve the quality of service. Patients can avoid waiting lines and doctors don't waste time for unnecessary consultations and paperwork. Telemedicine also improves the availability of care as patients' state can be monitored and consulted anywhere and anytime.

9. Integrating Big Data With Medical Imaging

Medical imaging is vital and each year in the US about 600 million imaging procedures are performed. Analyzing and storing manually these images is expensive both in terms of time and money, as radiologists need to examine each image individually, while hospitals need to store them for several years.

Medical imaging provider Carestream explains how big data analytics for healthcare could change the way images are read: algorithms developed analyzing hundreds of thousands of images could identify specific patterns in the pixels and convert it into a number to help the physician with the diagnosis. They even go further, saying that it could be possible that radiologists will no longer need to look at the images, but instead analyze the outcomes of the algorithms that will inevitably study and remember more images than they could in a lifetime. This would undoubtedly impact the role of radiologists, their education and required skillset.

Real world applications of big data in healthcare

Organizations have been generating data for decades; even now large amount of data is being generated daily. Big data is a term often used to represent such data. Well, there is no particular definition for it, but usually it refers to large amount of data which may be structured or unstructured and comes from different sources. For efficient business operations and profits, big data is analysed carefully by the organizations to reach better decisions.

Currently, this technology is being used in wide range of areas but one of the areas where it can bring a hugechange is **healthcare**.

Need of Big data in Healthcare

Of course, it comes to our mind that why there is a need of big data in healthcare systems, well there are some reasons-

The physicians now-a-days rely more on patient's clinical health record which means gathering of large amount of data, that too for different patients. Surely, this cannot be easily done with old techniques of storing the data.

There is large amount of data coming in from healthcare systems either from billing systems or from EMR (Electronic Medical Records). There is certainly large variety of data coming from different sources, in different formats driving the need for big data approach to tackle all this.

Challenges towards data-driven Healthcare

Health systems today generate lots of data from different sources such as laboratory tests, clinical notes, patient's reports, etc. The real challenge is how to collect, analyse and manage such huge information to predict the outcomes and make possible decisions.

Medical data today is spread across different sources governed by different hospitals and departments. So, there is a need for the development of new infrastructure which can integrate all the data from such sources.

Real Life Examples:

1. Predictive Analytics in Healthcare

Everyone is a patient at one time or the other and all need good medical care. We believe doctors are medical experts and what they decide for us is best. But ever thought how difficult it would have been for them to analyse the patient's entire history and make proper decisions for their treatments?

Predictive analysis leads to patient's safety and quality care. It keeps doctors informed about the patient's medical histories and helps predict results for future. For example, the analytics tools would be able to predict which patient is at risk of what disease, so to make decisions accordingly to improve patient's health. Predictive algorithms using different programming languages can be created to predict the health of a patient over time.

To improve the healthcare systems, a US Research collaborative, Optum Labs collected data of over 30 million patients to create a database for predictive analytics tools that will improve healthcare systems.

2. Electronic Health Records (EHRs)

The volume and details of patient's record is increasing rapidly and there arises the need of adopting a new approach. Many hospitals have moved over to use **Electronic Health Records (EHRs)** which is the main application of big data in healthcare. Every patient has his/her own medical records such as laboratory tests results, medical reports, lists of medicines, etc. EHRs make it easier to maintain the data and have access to such data.

A separate file or record is maintained of each patient that can be easily modified time to time by the doctor and these records can be shared safely.

3. Real-Time Monitoring

Healthcare Systems are looking forward to offer better treatments to their patients by constantly monitoring their health in real-time. Many tools are there which analyse the data of the patient and advice the doctors to take respective actions. For example, **new wearable sensors** can help track patient's health trends that can be monitored by the doctors. They can be helpful from tracking blood pressure to other illnesses right at home, which in turn will reduce patient's unnecessary visits to the clinics.

4. Prevention of Unnecessary ER visits

Hospitals want to reduce the number of ER visits or Emergency visits of patients. They believe that it increases healthcare costs and sometimes does not lead to better outcomes for patients. For example, a man suffering with acute abdominal pain comes to an emergency room. The doctor will try to figure out the cause of the problem such as kidney stone or appendicitis or something else. Now if he has a way of knowing the patient's past medical results, he could begin the treatment as soon as possible. The examination would take less time and would also cost less money.

For this, **Almeda county hospitals** in California, USA planned to create a program which called **PreManage** ED. According to this program, the records of the patients are shared with the emergency departments such as, if the patient has already done some tests at other hospitals or earlier what advice were given to the patient. This reduces the time of patient to get the details of previous tests to them and do unnecessary formalities. This is indeed a great application of big data analytics in healthcare area which saves both time and money.

5. Big data can help cure cancer

Cancer is a complex disease where a single tumour can have billions of cells. Hearing this word, we think that it can be cured only at hospitals and not at computer rooms. Well, **medical researchers can use analytics to see the recovery rates of cancer patients and the treatment plans to find the treatments that have highest rates of success for this disease**. To make this successful, patient's database from different health institutions need to be linked up keeping in mind confidentiality of patient's data.

For example, patient's tumor samples can be examined with their other treatment records which in turn, will help researchers to carry the treatment accordingly. Finding such trends will lead to better results. This approach is not just limited to cancer but can be used to other diseases as well.

These are the few ways in which big data analytics is having impact on healthcare. With the use of advanced big data analytics, healthcare providers can help improve patient outcomes, while lowering the costs at the same time. If you are a startup enthusiast in this space, these are some core areas where new services can be offered.

Use of Big Data Analytics in Healthcare

Healthcare analytics have the possible to cut down amounts of hospitalization, anticipate outbreaks of infectious, evade avoidable epidemics and boost the nature of activity in general. The regular human life period is developing along world community, which acts as new objections to present hospitalization transmission approach.

- Predict the daily patients income to tailor staffing accordingly
- Use Electronic Health Records (EHRs)
- Use real-time alerting for instant care
- Help in preventing opioid (drugs) abuse in the US
- Enhance patient engagement in their own health
- Use health data for a better-informed strategic planning
- Research more extensively to cure cancer
- Use predictive analytics
- Reduce fraud and enhance data security
- Practice telemedicine
- Integrate medical imaging for an broader diagnosis
- Prevent unnecessary ER visits
- Heart Attack Prediction
- Brain Disease Prediction
- Prediction of Disease Outcome
- Using Hive & R Analyzing Diabetics
- Analysis of Coronary Artery Heart Disease
- Infectious Disease Outbreak Prediction
- Tuberculosis Prediction.
- Early Stage Heart Attack Detection
- Intelligent Heart Disease Prediction System
- Diagnose of Chronic Kidney Disease
- HIV/AIDS Disease Prediction

1.7 Big Data in Advertising

The digital advertising industry is evolving like never before. The ability to capture and analyze massive amounts of structured and unstructured data is helping digital advertisers to discover new relationships, spot emerging trends and patterns, and gain actionable insights that lead to competitive advantage. As a result, traditional advertising is shifting rapidly into the realm of personalized and highly targeted online and mobile ads—the realm of data driven marketing. Once dismissed as a “buzzword”, big data is having a big impact on the digital advertising industry, and here are some reasons why.

Finding order among chaos

Successful digital advertising depends upon the ability to collect, integrate and analyze data from both internal and external sources. The challenge lies in the fact that 80% of that data is unstructured or "chaotic". This is data from sources such as photos, videos and social media posts—data that says so much about us—but cannot be analyzed via traditional methods. By using big data analytics platforms, companies are now able to capture, store and analyze all collected data, both structured and unstructured. As a result, digital advertisers can gain fresh and relevant insights from raw chaotic data. Actionable insights that inform marketing decisions and strategies.

Real-time data analysis

In the past, conventional scalable relational database solutions could be relied upon to effectively manage and analyze massively large data sets. But they did so at a snail-like pace, taking days and even weeks to perform tasks that often yielded "stale" results. By contrast, the big data analytics platforms of today can perform sophisticated processes at lightning-fast speeds, allowing for real-time analysis and insights. Shorter time to insight allows marketers to make real-time decisions and take immediate action based on fresh, reliable and relevant information.

More personalized and targeted ads

Big data allows digital advertisers to better target users with more personalized ads that they most likely want to see. Google, and now Facebook—the dominant players in digital advertising—have gotten very good at creating and delivering more appealing ads in non-intrusive ways. Ads featuring products and services we might actually want and use to better our lives. And these more personalized and targeted ads are all based on massive amounts of personal data we constantly provide about what we're doing, saying, liking, sharing—and now thanks to our mobile devices—where we're going. Which brings us to…

Hyper-localized advertising

The proliferation of mobile devices, primarily smartphones, has created a major opportunity for digital advertisers to deliver mobile specific ads to the right people at the right time—in context. Through the combination of social data and location data, stores that shoppers are near and might be interested in can send out ads offering percentage discounts or other incentives—delivered to the shopper's location in real time—to get them to walk through their doors. Hyper-localized advertising has been shown to increase customer engagement and conversion rates. However, there is potential for backlash as some customers may get a creepy feeling upon realizing that advertisers actually know where they are in real-time. As a result, advertisers will need to make some tradeoffs in order to keeps their ads effective while mitigating offenses.

Mergers and acquisitions

In the world of digital based advertising—a world primarily dominated by Google and Facebook—more corporate mergers and acquisitions will need to take place in order for companies to gain the economies of scale needed to compete against the giants. It was for that very reason that the recent merger of the French advertising company Publicis with the American advertising company Omnicom took place. The merger serves as a cautionary tale for advertisers as they go digital. Going forward, only those companies that have the talent, tools and infrastructure needed to compete in the high-stakes digital advertising industry will survive.

While digital advertising comprises about 25% of today's total advertising spend, the full impact of big data will be felt in the future when all advertising will be data driven.

1. Audience Predictions

As publishers and media companies begin their data-driven journeys, for the first time data is being used on a large scale in order to deliver the right content to the right people on the right platform at the right time.
With the scope of big data being collected nowadays and the potential to mine it to understand what content, movies and music consumers want is huge. It has been highlighted that data obtained on user behaviors via social media often reveals overlooked factors that have the potential to drive customer interest.

As consumers nowadays have the choice to select from formats such as on-demand, streaming media, pay-per-view, subscription-based and many more, most content is now delivered via various digital channels allowing media houses to collect, analyze and interpret user data efficiently and effectively. For example, Netflix interpreted the amount of viewership information to conclude which fiction dramas were favored by their consumers. With this data, it secured the right to broadcast the most favored drama 'Houses of Cards' by outbidding competitors.

Similarly, YouTube has interpreted vital statistics in order to deliver users with what they like the most. Information gathered allowed YouTube to learn about what video viewers enjoyed the most, which devices were used for streaming and the duration for which particular videos were viewed.

2. Improved Targeted Advertising

With advertising crucial in the media industry, in previous years it was purely conducted on the basis of assumptions. However, nowadays companies are harnessing big data in order to understand preferences providing in-depth customer insights such as when they would watch advertisements and at which particular time. This improved visibility helps ad ops position advertisements at specific time slots for higher conversion rates.

As big data makes it possible to understand digital media consumption, behavior can be used with traditional demographic data to provide personalized advertising. Big data applications improve ad targeting amid increasingly complex content consumption behavior. As consumers nowadays have access to multiple devices, big data insights help to understand when consumers use a second screen so that campaigns can be optimized across devices. Digital conversion rates can also be increased by offering micro-segmentation of customers across advertising networks and exchanges.

With social media platforms and YouTube providing better data for targeted advertising, it cannot be denied that TV still attracts attention. According to Fox Media, in 2017 they confirmed that digital viewers and TV viewings tuning their content were able to see the same advertisements. These advertisements were selected on the basis of Video Quality Score (VQS), using Moat which provides real-time, multi-platform and actionable marketing analytics. In addition to Fox, leading media companies such as NBC, Vice, The New York Times and CBC have begun using Moat.

Due to the preceding benefits, big data analytics is slowly becoming a choice for various media organizations worldwide. It creates an ecosystem allowing customers to take center stage. Ultimately, success in the media industry entirely depends on the user-experiences which deliver.

3. Expanded Customer Acquisition & Retention

Increasing customer churn is highly important for media companies. Nowadays, most customers resort to both social media and review sites before viewing particular series, movies, shows, music programs or downloading publications. The evolution of big data has allowed media companies to design tailored strategies in order to attract and retain customers. By leveraging various data sets, companies can understand consumer's likes and dislikes, why consumers subscribe and unsubscribe thus helping media companies develop and tailor attractive promotional and product strategies and in order to attract and retain customers.

Unstructured big data sources are best handled by data applications such as call detail records, email, and social media can often be overlooked factors when looking at customer interests and churn.

Companies such as Warner Bros implemented software applications with sales data in order to gain quick access to actionable, accurate reports in order to support and accumulate knowledge in order to obtain insights to expand customer acquisition and retention.

4. New Product Development & Content Monetization

Big data and analytics can aid media organizations to generate additional sources of revenue. With accurate data, incentivization of consumer behavior can be undertaken which can help reveal the true market value of content or identify potential new product or service opportunities.

An example of this includes the Weather Company, owners of The Weather Channel (TWC) which is also co-owned by IBM. TWC used big data in order to observe and under customer behavior with regards to specific weather conditions.
Using the data available to them, they have created the new WeatherFX marketplace which allows advertisers to correlate their display advertisements with weather events based on various products which would most likely to sell in correlation with particular weather conditions. TWC is estimated to earn at least half of their advertising revenue using the results of big data analytics.

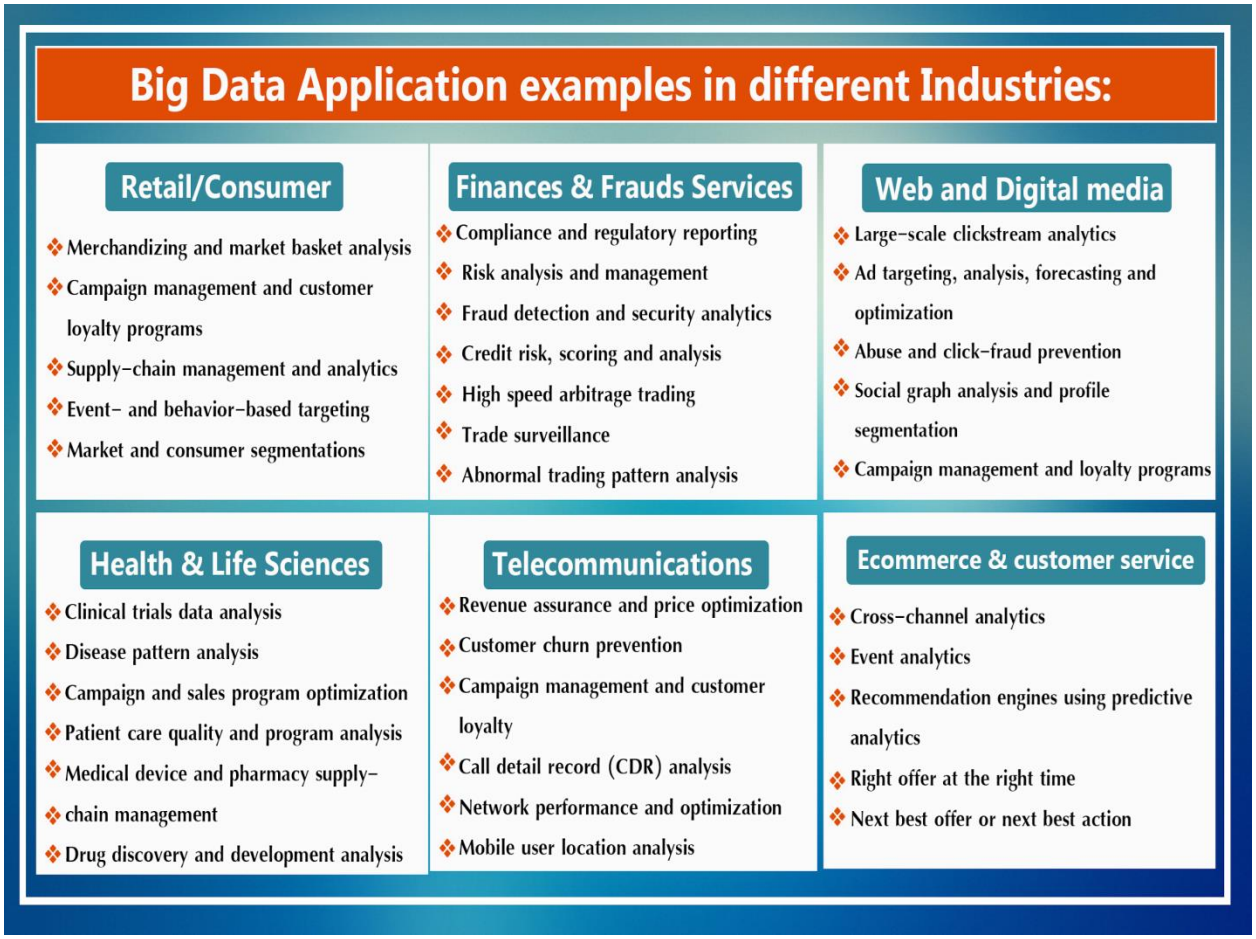
Mobile profusion and bandwidth expansion make it possible to engage with a larger chunk of digitally connected audiences for content monetization as big data facilitates zoning in the right content which the audience prefers.

5. Media Scheduling Optimization

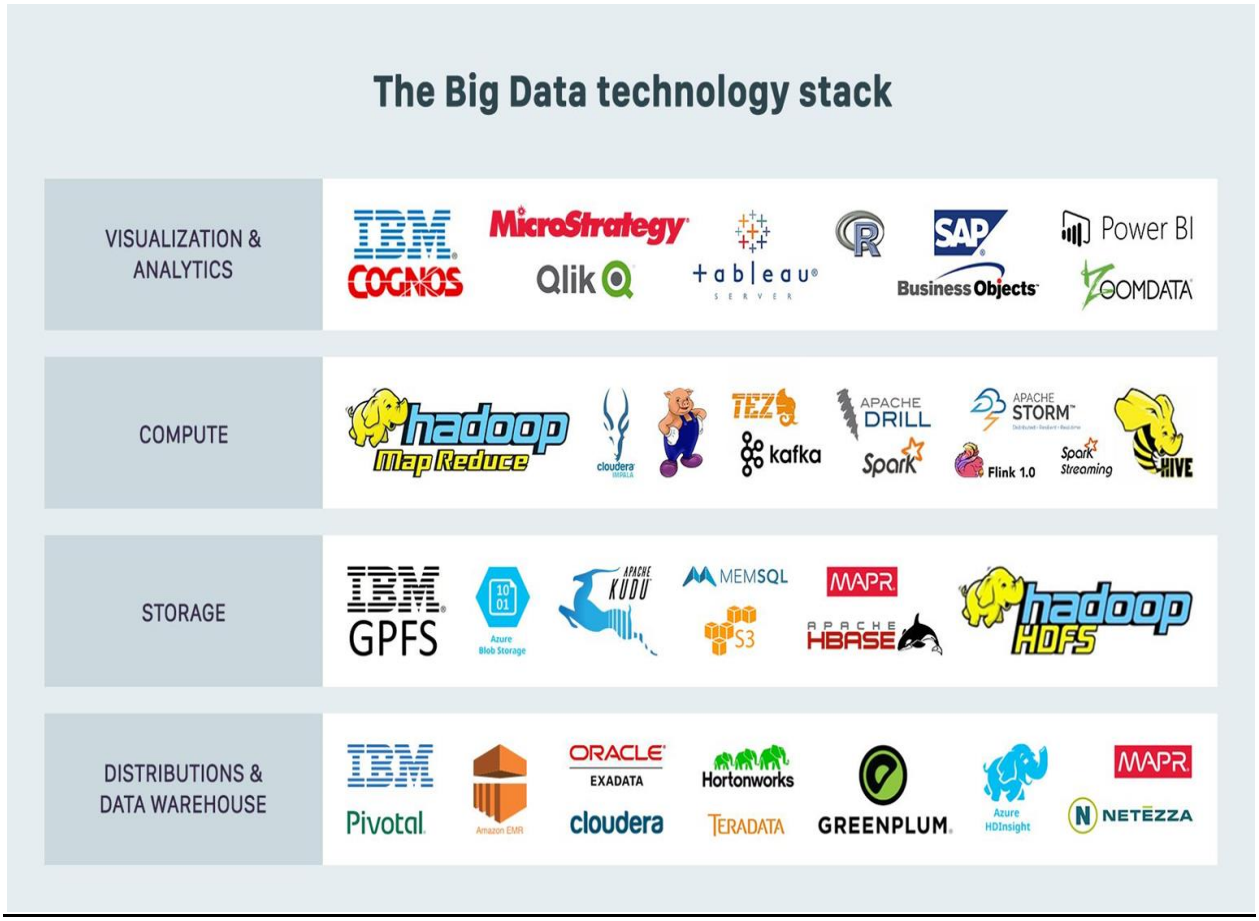
Up until a couple of years ago, there was a gap between both the distributors and consumers. However, with the evolution of digital media platforms, this has changed, making it easier for distributors to approach potential consumers without the need for an intermediary.
Social networks allow distributors to create personal connections with consumers. Connecting with consumers via the scheduling of media streams in order to maximize profits. With the

scaling ability of big data, information can be analyzed at granular levels such as ZIP code levels for localized distribution.

An example of this includes the release of Bollywood film Chennai Express, prior to the release, movie-related tweets generated more than 1 billion impressions and tweets with relevant hashtags generated more than 750,000 impressions during a 3-month campaign. As a result of a number of data analytics sources and marketing the film established several new records which to date remain unchallenged. It is probably one of the quickest films to enter the billion-dollar club.



1.8 Big Data Technologies:



15 Big Data Technologies to Watch

The list of technology vendors offering big data solutions is seemingly infinite. Many of the big data solutions that are particularly popular right now fit into one of the following 15 categories:

1. The Hadoop Ecosystem

While Apache Hadoop may not be as dominant as it once was, it's nearly impossible to talk about big data without mentioning this open source framework for distributed processing of large data sets. Last year, Forrester predicted, "100% of all large enterprises will adopt it (Hadoop and related technologies such as Spark) for big data analytics within the next two years."

Over the years, Hadoop has grown to encompass an entire ecosystem of related software, and many commercial big data solutions are based on Hadoop. In fact, Zion Market Research forecasts that the market for Hadoop-based products and services will continue to grow at a 50 percent CAGR through 2022, when it will be worth $87.14 billion, up from $7.69 billion in 2016.

Key Hadoop vendors include Cloudera, Hortonworks and MapR, and the leading public clouds all offer services that support the technology.

2. Spark

Apache Spark is part of the Hadoop ecosystem, but its use has become so widespread that it deserves a category of its own. It is an engine for processing big data within Hadoop, and it's up to one hundred times faster than the standard Hadoop engine, MapReduce.

In the AtScale 2016 Big Data Maturity Survey, 25 percent of respondents said that they had already deployed Spark in production, and 33 percent more had Spark projects in development. Clearly, interest in the technology is sizable and growing, and many vendors with Hadoop offerings also offer Spark-based products.

3. R

R, another open source project, is a programming language and software environment designed for working with statistics. The darling of data scientists, it is managed by the R Foundation and available under the GPL 2 license. Many popular integrated development environments (IDEs), including Eclipse and Visual Studio, support the language.

Several organizations that rank the popularity of various programming languages say that R has become one of the most popular languages in the world. For example, the IEEE says that R is the fifth most popular programming language, and both Tiobe and RedMonk rank it 14th. This is significant because the programming languages near the top of these charts are usually general-purpose languages that can be used for many different kinds of work. For a language that is used almost exclusively for big data projects to be so near the top demonstrates the significance of big data and the importance of this language in its field.

4. Data Lakes

To make it easier to access their vast stores of data, many enterprises are setting up data lakes. These are huge data repositories that collect data from many different sources and store it in its natural state. This is different than a data warehouse, which also collects data from disparate sources, but processes it and structures it for storage. In this case, the lake and warehouse metaphors are fairly accurate. If data is like water, a data lake is natural and unfiltered like a body of water, while a data warehouse is more like a collection of water bottles stored on shelves.

Data lakes are particularly attractive when enterprises want to store data but aren't yet sure how they might use it. A lot of Internet of Things (IoT) data might fit into that category, and the IoT trend is playing into the growth of data lakes.

MarketsandMarkets predicts that data lake revenue will grow from $2.53 billion in 2016 to $8.81 billion by 2021.

5. NoSQL Databases

Traditional relational database management systems (RDBMSes) store information in structured, defined columns and rows. Developers and database administrators query, manipulate and manage the data in those RDBMSes using a special language known as SQL.

NoSQL databases specialize in storing unstructured data and providing fast performance, although they don't provide the same level of consistency as RDBMSes. Popular NoSQL databases include MongoDB, Redis, Cassandra, Couchbase and many others; even the leading RDBMS vendors like Oracle and IBM now also offer NoSQL databases.

NoSQL databases have become increasingly popular as the big data trend has grown. According to Allied Market Research the NoSQL market could be worth $4.2 billion by 2020. However, the market for RDBMSes is still much, much larger than the market for NoSQL.

6. Predictive Analytics

Predictive analytics is a sub-set of big data analytics that attempts to forecast future events or behavior based on historical data. It draws on data mining, modeling and machine learning techniques to predict what will happen next. It is often used for fraud detection, credit scoring, marketing, finance and business analysis purposes.

In recent years, advances in artificial intelligence have enabled vast improvements in the capabilities of predictive analytics solutions. As a result, enterprises have begun to invest more in big data solutions with predictive capabilities. Many vendors, including Microsoft, IBM, SAP, SAS, Statistica, RapidMiner, KNIME and others, offer predictive analytics solutions. Zion Market Research says the Predictive Analytics market generated $3.49 billion in revenue in 2016, a number that could reach $10.95 billion by 2022.

7. In-Memory Databases

In any computer system, the memory, also known as the RAM, is orders of magnitude faster than the long-term storage. If a big data analytics solution can process data that is stored in memory, rather than data stored on a hard drive, it can perform dramatically faster. And that's exactly what in-memory database technology does.

Many of the leading enterprise software vendors, including SAP, Oracle, Microsoft and IBM, now offer in-memory database technology. In addition, several smaller companies like Teradata, Tableau, Volt DB and DataStax offer in-memory database solutions. Research from MarketsandMarkets estimates that total sales of in-memory technology were $2.72 billion in 2016 and may grow to $6.58 billion by 2021.

8. Big Data Security Solutions

Because big data repositories present an attractive target to hackers and advanced persistent threats, big data security is a large and growing concern for enterprises. In the AtScale survey, security was the second fastest-growing area of concern related to big data.

According to the IDG report, the most popular types of big data security solutions include identity and access controls (used by 59 percent of respondents), data encryption (52 percent) and data segregation (42 percent). Dozens of vendors offer big data security solutions, and Apache Ranger, an open source project from the Hadoop ecosystem, is also attracting growing attention.

9. Big Data Governance Solutions

Closely related to the idea of security is the concept of governance. Data governance is a broad topic that encompasses all the processes related to the availability, usability and integrity of data. It provides the basis for making sure that the data used for big data analytics is accurate and appropriate, as well as providing an audit trail so that business analysts or executives can see where data originated.

In the NewVantage Partners survey, 91.8 percent of the Fortune 1000 executives surveyed said that governance was either critically important (52.5 percent) or important (39.3 percent) to their big data initiatives. Vendors offering big data governance tools include Collibra, IBM, SAS, Informatica, Adaptive and SAP.

10. Self-Service Capabilities

With data scientists and other big data experts in short supply — and commanding large salaries — many organizations are looking for big data analytics tools that allow business users to self-service their own needs. In fact, a report from Research and Markets estimates that the self-service business intelligence market generated $3.61 billion in revenue in 2016 and could grow to $7.31 billion by 2021. And Gartner has noted, "The modern BI and analytics platform emerged in the last few years to meet new organizational requirements for accessibility, agility and deeper analytical insight, shifting the market from IT-led, system-of-record reporting to business-led, agile analytics including self-service."

Hoping to take advantage of this trend, multiple business intelligence and big data analytics vendors, such as Tableau, Microsoft, IBM, SAP, Splunk, Syncsort, SAS, TIBCO, Oracle and other have added self-service capabilities to their solutions. Time will tell whether any or all of the products turn out to be truly usable by non-experts and whether they will provide the business value organizations are hoping to achieve with their big data initiatives.

11. Artificial Intelligence

While the concept of artificial intelligence (AI) has been around nearly as long as there have been computers, the technology has only become truly usable within the past couple of years. In many ways, the big data trend has driven advances in AI, particularly in two subsets of the discipline: machine learning and deep learning.

The standard definition of machine learning is that it is technology that gives "computers the ability to learn without being explicitly programmed." In big data analytics, machine learning technology allows systems to look at historical data, recognize patterns, build models and predict future outcomes. It is also closely associated with predictive analytics.

Deep learning is a type of machine learning technology that relies on artificial neural networks and uses multiple layers of algorithms to analyze data. As a field, it holds a lot of promise for allowing analytics tools to recognize the content in images and videos and then process it accordingly.

Experts say this area of big data tools seems poised for a dramatic takeoff. IDC has predicted, "By 2018, 75 percent of enterprise and ISV development will include cognitive/AI or machine learning functionality in at least one application, including all business analytics tools."

Leading AI vendors with tools related to big data include Google, IBM, Microsoft and Amazon Web Services, and dozens of small startups are developing AI technology (and getting acquired by the larger technology vendors).

12. Streaming analytics

As organizations have become more familiar with the capabilities of big data analytics solutions, they have begun demanding faster and faster access to insights. For these enterprises, streaming analytics with the ability to analyze data as it is being created, is something of a holy grail. They are looking for solutions that can accept input from multiple disparate sources, process it and return insights immediately — or as close to it as possible. This is particular desirable when it comes to new IoT deployments, which are helping to drive the interest in streaming big data analytics.

Several vendors offer products that promise streaming analytics capabilities. They include IBM, Software AG, SAP, TIBCO, Oracle, DataTorrent, SQLstream, Cisco, Informatica and others. MarketsandMarkets believes the streaming analytics solutions brought in $3.08 billion in revenue in 2016, which could increase to $13.70 billion by 2021.

13. Edge Computing

In addition to spurring interest in streaming analytics, the IoT trend is also generating interest in edge computing. In some ways, edge computing is the opposite of cloud computing. Instead of transmitting data to a centralized server for analysis, edge computing systems analyze data very close to where it was created — at the edge of the network.

The advantage of an edge computing system is that it reduces the amount of information that must be transmitted over the network, thus reducing network traffic and related costs. It also decreases demands on data centers or cloud computing facilities, freeing up capacity for other workloads and eliminating a potential single point of failure.

While the market for edge computing, and more specifically for edge computing analytics, is still developing, some analysts and venture capitalists have begun calling the technology the "next big thing."

14. Blockchain

Also a favorite with forward-looking analysts and venture capitalists, blockchain is the distributed database technology that underlies Bitcoin digital currency. The unique feature of a blockchain database is that once data has been written, it cannot be deleted or changed after the fact. In addition, it is highly secure, which makes it an excellent choice for big data applications in sensitive industries like banking, insurance, health care, retail and others.

Blockchain technology is still in its infancy and use cases are still developing. However, several vendors, including IBM, AWS, Microsoft and multiple startups, have rolled out experimental or introductory solutions built on blockchain technology.

15. Prescriptive Analytics

Many analysts divide big data analytics tools into four big categories. The first, descriptive analytics, simply tells what happened. The next type, diagnostic analytics, goes a step further and provides a reason for why events occurred. The third type, predictive analytics, discussed in depth above, attempts to determine what will happen next. This is as sophisticated as most analytics tools currently on the market can get.

However, there is a fourth type of analytics that is even more sophisticated, although very few products with these capabilities are available at this time. Prescriptive analytics offers advice to companies about what they should do in order to make a desired result happen. For example, while predictive analytics might give a company a warning that the market for a particular product line is about to decrease, prescriptive analytics will analyze various courses of action in response to those market changes and forecast the most likely results.

Currently, very few enterprises have invested in prescriptive analytics, but many analysts believe this will be the next big area of investment after organizations begin experiencing the benefits of predictive analytics.

The market for big data technologies is diverse and constantly changing. But perhaps one day soon predictive and prescriptive analytics tools will offer advice about what is coming next for big data — and what enterprises should do about it.

16.Apache Hadoop

Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

17. Microsoft HDInsight

It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability with low cost.

18. NoSQL

While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL (Not Only SQL) to handle unstructured data. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs available to analyse big Data.

19. Hive

This is a distributed data management for Hadoop. This supports SQL-like query option HiveSQL (HSQL) to access big data. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

20. Sqoop

This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

21. PolyBase

This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a datawarhousing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

22. Big data in EXCEL

As many people are comfortable in doing analysis in EXCEL, a popular tool from Microsoft, you can also connect data stored in Hadoop using EXCEL 2013. Hortonworks, which is primarily working in providing Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. You can use Power View feature of EXCEL 2013 to easily summarise the data.

Similarly, Microsoft's HDInsight allows us to connect to Big data stored in Azure cloud using a power query option.

23. Presto

Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.