NAGARJUNA COLLEGE OF ENGINEERING AND TECHOLOGY
DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Big Data

(17ISI732)

Module 1

Overview of Big Data

By,

Mrs. Monika N

Asst.

Professor ISE,

NCET

Module-1:Overview of Big Data

Introduction

Big data is a term defined for data sets that are large or complex that traditional data processing applications are inadequate. Big Data basically consists of analysis zing, capturing the data, data creation, searching, sharing, storage capacity, transfer, visualization, and querying and information privacy.

What is Big Data?

- ✔ **Big Data** is a collection of large datasets that cannot be adequately processed using traditional processing techniques. Big data is not only data it has become a complete subject, which involves various tools, techniques and frameworks.
- ✔ Big data term describes the volume amount of data both structured and unstructured manner that adapted in day-to-day business environment. It's important that what organizations utilize with these with the data that matters.
- ✔ Big data helps to analyze the in-depth concepts for the better decisions and strategic taken for the development of the organization.

Types Of Big Data

Big Data could be found in three forms:

1. Structured
2. Unstructured
3. Semi-structured

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. 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.

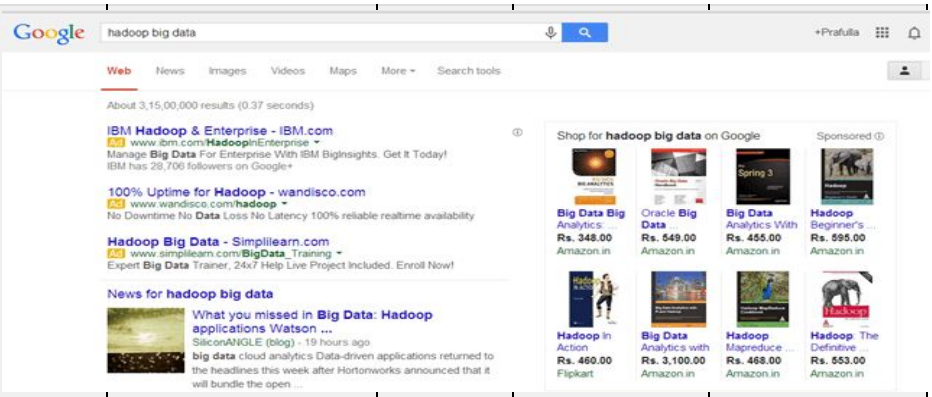
Examples Of Structured Data:

An

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

Unstructured

- ✔ Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. 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



About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/HadoopInEnterprise
Manage Big Data For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/hadoop
No Downtime No Data Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/BigData_Training
Expert Big Data Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for hadoop big data on Google Sponsored

Big Data Big Analytics: ... Rs. 348.00 Amazon.in	Oracle Big Data ... Rs. 549.00 Amazon.in	Big Data Analytics With Rs. 455.00 Amazon.in	Hadoop Beginner's ... Rs. 595.00 Amazon.in
Hadoop In Action Rs. 460.00 Flipkart	Big Data Analytics with Rs. 3,100.00 Amazon.in	Hadoop Mapreduce ... Rs. 468.00 Amazon.in	Hadoop: The Definitive ... Rs. 553.00 Amazon.in

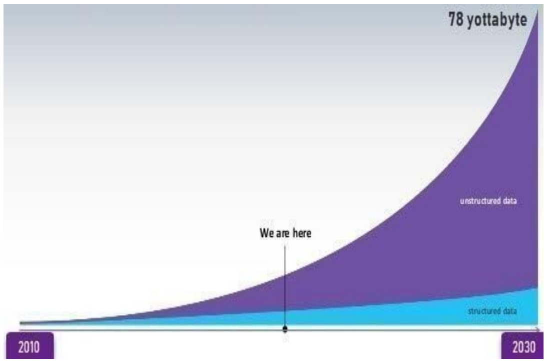
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

- ✔ The output returned by 'Google Search'

Semi-structured

- Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.
- Examples Of Semi-structured Data
- Personal data stored in an XML file-

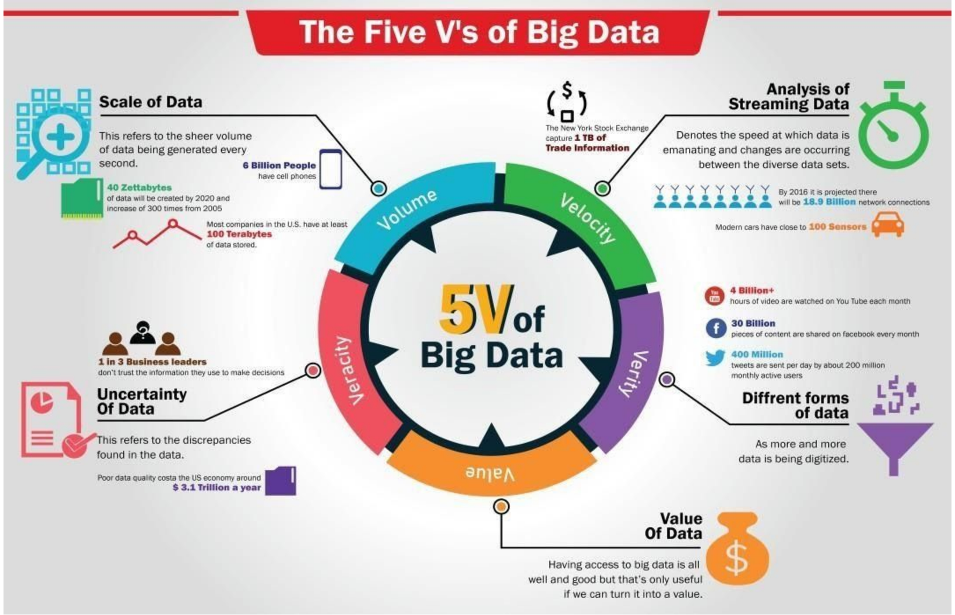
- <rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
- <rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
- <rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
- <rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
- <rec><name>Jeremiah J.</name><sex>Male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



- Please note that web application data, which is unstructured, consists of log files, transaction history files etc. OLTP systems are built to work with structured data wherein data is stored in relations (tables).

The Evolution of Big Data

While the term “big data” is the new in this era, as it is the act of gathering and storing huge amounts of information for eventual analysis is ages old. The concept came into existence in the early 2000s when Industry analyst Doug Laney the definition of big data as the three categories as follows:


The Five V's of Big Data
Scale of Data
This refers to the sheer volume of data being generated every second.
6 Billion People have cell phones
40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005
Most companies in the U.S. have at least 100 Terabytes of data stored.
The New York Stock Exchange capture 1 TB of Trade Information
Analysis of Streaming Data
Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.
By 2016 it is projected there will be 18.9 Billion network connections
Modern cars have close to 100 Sensors
Volume
Velocity
Verity
Value
Veracity
5V of Big Data
4 Billion+ hours of video are watched on You Tube each month
30 Billion pieces of content are shared on facebook every month
400 Million tweets are sent per day by about 200 million monthly active users
Diffrent forms of data
As more and more data is being digitized.
1 in 3 Business leaders don't trust the information they use to make decisions
Uncertainty Of Data
This refers to the discrepancies found in the data.
Poor data quality costa the US economy around $ 3.1 Trillion a year
Value Of Data
Having access to big data is all well and good but that's only useful if we can turn it into a value.

Volume: Organizations collects the data from relative sources, which includes business transactions, social media and information from sensor or machine-to-machine data. Before, storage was a big issue but now the advancement of new technologies (such as Hadoop) has reduce the burden.

Velocity: Data streams unparalleled speed of velocity and have improved in timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in real time operations.

Variety: Data comes in all varieties in form of structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

In SAS, we consider two additional dimensions with respect to big data:

Velocity

First let's talk about velocity. Obviously, velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every day the number of emails, twitter messages, photos, video clips, etc. increases at lighting speeds around the world. Every second of every day data is increasing. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access to website, credit card verification and instant messaging. Big data technology allows us now to analyze the data while it is being generated, without ever putting it into databases.

Volume

Volume refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc. The vast amounts of data have become so large in fact that we can no longer store and analyze data using traditional database technology. We now use distributed systems, where parts of the data is stored in different locations and brought together by software. With just Facebook alone there are 10 billion messages, 4.5

billion times that the "like" button is pressed, and over 350 million new pictures are uploaded every day. Collecting and analyzing this data is clearly an engineering challenge of immensely vast proportions.

Value

When we talk about value, we're referring to the worth of the data being extracted. Having endless amounts of data is one thing, but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data. The most important part of embarking on a big data initiative is to understand the costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is reaped can be monetized.

Variety

Variety is defined as the different types of data we can now use. Data today looks very different than data from the past. We no longer just have structured data (name, phone number, address, financials, etc) that fits nice and neatly into a data table. Today's data is unstructured. In fact, 80% of all the world's data fits into this category, including photos, video sequences, social media updates, etc. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

Veracity

Last, but certainly not least there is veracity. Veracity is the quality or trustworthiness of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content. Gleaning loads and loads of data is of no use if the quality or trustworthiness is not accurate. Another good example of this relates to the use of GPS data. Often the GPS will "drift" off course as you peruse through an urban area. Satellite signals are lost as they bounce off tall buildings or other structures. When this happens, location data has to be fused with another data source like road data, or data from an accelerometer to provide accurate data.

Ignoring Big Data won't make it go away, and while it may not immediately kill your business it shouldn't be ignored for very long. The results of Big Data can generally be directly measured making it easy to determine a return on investment. Big Data is a tool definitely worth looking into.

What are the categories which come under Big Data?

Big data works on the data produced by various devices and their applications. Below are some of the fields that are involved in the umbrella of Big Data.

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.

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.

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.

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

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

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

What is the importance of Big Data?

The importance of big data is how you utilize the data which you own. Data can be fetched from any source and analyze it to solve that enable us in terms of

- Cost reductions
- Time reductions,
- New product development and optimized offerings, and
- Smart decision making.

Combination of big data with high-powered analytics, you can have great impact on your business strategy such as:

- Finding the root cause of failures, issues and defects in real time operations.
- Generating coupons at the point of sale seeing the customer's habit of buying goods.
- Recalculating entire risk portfolios in just minutes.
- Detecting fraudulent behavior before it affects and risks your organization.

Who are the ones who use the Big Data Technology?

Banking

Large amounts of data streaming in from countless sources, banks have to find out unique and innovative ways to manage big data. It's important to analyze customers needs and provide them service as per their requirements, and minimize risk and fraud while maintaining regulatory compliance. Big data have to deal with financial institutions to do one step from the advanced analytics.

Government

When government agencies are harnessing and applying analytics to their big data, they have improvised a lot in terms of managing utilities, running agencies, dealing with traffic congestion or preventing the affects crime. But apart from its advantages in Big Data, governments also address issues of transparency and privacy.

Education

Educator regarding Big Data provides a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, ensuring student's progress, and can implement an improvised system for evaluation and support of teachers and principals in their teachings.

Health Care

When it comes to health care in terms of Patient records. Treatment plans. Prescription information etc., everything needs to be done quickly and accurately and some aspects enough transparency to satisfy stringent industry regulations. Effective management results in good health care to uncover hidden insights that improve patient care.

Manufacturing

Manufacturers can improve their quality and output while minimizing waste where processes are known as the main key factors in today's highly competitive market. Several manufacturers are working on analytics where they can solve problems faster and make more agile business decisions.

Retail

Customer relationship maintains is the biggest challenge in the retail industry and the best way to manage will be to manage big data. Retailers must have unique marketing ideas to sell their products to customers, the most effective way to handle transactions, and applying improvised tactics of using innovative ideas using Big Data to improve their business.

Brief explanation of how exactly businesses are utilizing Big Data?

Big Data is being converted into nuggets of information and then it becomes very straightforward for most business enterprises as we now know what their customers want, what are the products are rapidly fast moving, what are the expectations of the end users from the customer service, speed up the time sequence for marketing, methods on cost reduction, and methods to build economies of scale in a highly efficient manner. Hence Big Data leads to big time benefits for organizations and hence there exists a demand about it in the IT world.

Big Data Technologies

Accurate analysis carried out based on big data which helps to increase and optimizes operational efficiencies, enable cost reductions, and reduce risks for the business operations.

In order to capitalize on big data one should require infrastructure that manages and processes huge volumes of structured and unstructured data in real-time and can ensure data privacy and security.

Many technologies are available in the market from different vendors which includes Amazon, IBM, Microsoft, etc., to approach big data. To pick up a particular technology one must examine its classes, which areas are as follows

Operational Big Data

It includes the applications such as MongoDB which provides operational capabilities for interactive and real time workloads where data is generally captured and stored.

NoSQL Big Data systems are designed in such a way it capitalizes on new cloud computing architectures, to permit access on massive computations to be run reasonably and efficiently. Hence this builds operation on big data workloads much easier to manage, cheaper and faster to implement.

Analytical Big Data

- It owns the systems like Massively Parallel Processing database systems and MapReduce which provides the analytical capabilities for re collective and complex analysis.
- MapReduce provides a new method for analyzing the data that flaunts its capabilities provided by SQL, and based on a system called MapReduce that can be scaled up from single servers to thousands of high and low end machines.

Barriers

Barriers that are imposed on big data are as follows:

- Capture data
- Storage Capacity
- Searching
- Sharing
- Transfer
- Analysis
- Presentation

Enterprise servers are using the above measures to overcome the barriers mentioned above.

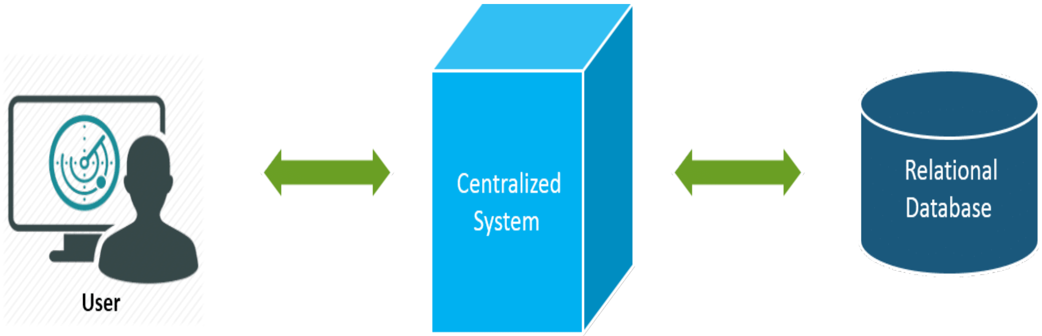
Differentiation between Operational vs. Analytical Systems

	Operational	**Analytical**
Latency	1 ms to 100 ms	1 min to 100 min

Concurrency	1000 to100,000	1 to 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL Database	MapReduce, MPP Database

Traditional Enterprise Approach

This approach of enterprise will use a computer to store and process big data. For storage purpose is available of their choice of database vendors such as Oracle, IBM, etc. The user interacts with the application, which executes data storage and analysis.



Limitation

This approach are good for those applications which require low storage, processing and database capabilities, but when it comes to dealing with large amounts of scalable data, it imposes a bottleneck.

Solution

Google solved this problem using an algorithm based on MapReduce. This algorithm divides the task into small parts or units and assigns them to multiple computers, and intermediate results together integrated results in the desired results.

Benefits of Big Data

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.

- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

DATA ANALYTICS

Data Science and Data Analytics are two most trending terminologies of today's time. Presently, data is more than oil to the industries. Data is collected into raw form and processed according to the requirement of a company and then this data is utilized for the decision making purpose. This process helps the businesses to grow & expand their operations in the market. But, the main question arises – What is the process called? Data Analytics is the answer here. And, Data Analyst and Data Scientist are the ones who perform this process.

WHAT IS DATA ANALYTICS?

Data or information is in raw format. The increase in size of the data has lead to a rise in need for carrying out inspection, data cleaning, transformation as well as data modeling to gain insights from the data in order to derive conclusions for better decision making process. This process is known as **data analysis**. Data Mining is a popular type of data analysis technique to carry out data modeling as well as knowledge discovery that is geared towards predictive purposes. Business Intelligence operations provide various data analysis capabilities that rely on data aggregation as well as focus on the domain expertise of businesses.

In Statistical applications, business analytics can be divided into **Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA)**. EDA focuses on discovering new features in the data and CDA focuses on confirming or falsifying existing hypotheses. Predictive Analytics does forecasting or classification by focusing on statistical or structural models while in text analytics, statistical, linguistic and structural techniques are applied to extract and classify information from textual sources, a species of unstructured data. All these are varieties of data analysis. The revolutionizing data wave has brought improvements to the overall functionalities in many different ways. There are various emerging requirements for applying advanced analytical techniques to the Big Data spectrum. Now experts can make more accurate and profitable decisions.

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.It is critical to design and built a data warehouse or Business Intelligence(BI) architecture that provides a flexible, multi-faceted analytical ecosystem, optimized for efficient ingestion and analysis of large and diverse data set. The analysis is an interactive process of a person tackling a problem, finding the data required to get an answer, analyzing that data, and interpreting the results in order to provide a recommendation for action.

A business intelligence environment, otherwise known as a reporting environment also includes calling as well as report execution. So, outputs are then printed in the desired form. Reporting refers to the process of organizing and summarizing data in an easily readable format to communicate important information. Reports help organizations in monitoring different areas of performance and improving customer satisfaction. One can also consider the conversion of raw data into useful information as a part of reporting, whereas, the same can be thought for analysis which transforms the information into key usable insights.

Difference between Data Analysis and Data Reporting

A report will show the user what had happened in the past to avoid inferences and help to get a feel for the data while analysis provides answers to any question or issue. An analysis process takes any steps needed to get the answers to those questions.

Reporting just provides the data that is asked for while analysis provides the information or the answer that is needed actually.

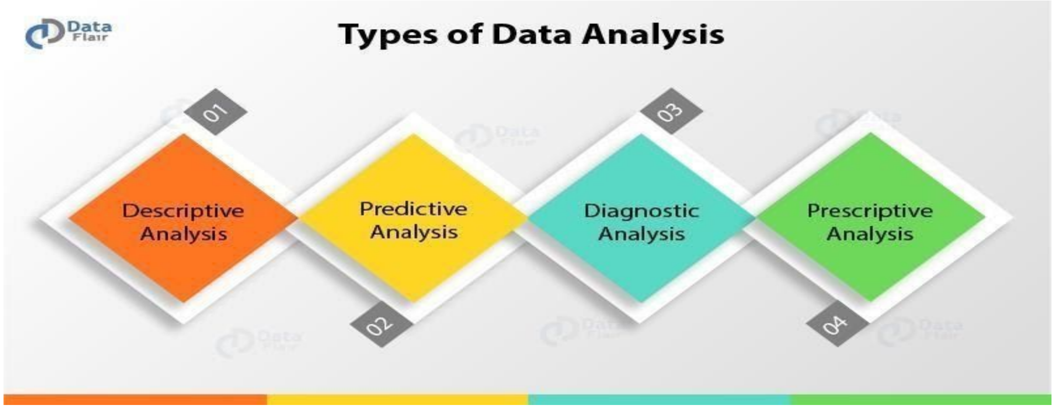
We perform the reporting in a standardized way, but we can customize the analysis. There are fixed standard formats for reporting while we perform the analysis as per the requirement; we customize it as needed.

We can perform reporting using a tool and it generally does not involve any person in the analysis. Whereas, a person is there for doing analysis and leading the complete analysis process.

Reporting is inflexible while analysis is flexible. Reporting provides no or limited context about what’s happening in the data and hence is inflexible while analysis emphasizes data points that are significant, unique, or special, and it explains why they are important to the business.

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

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.

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.

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.

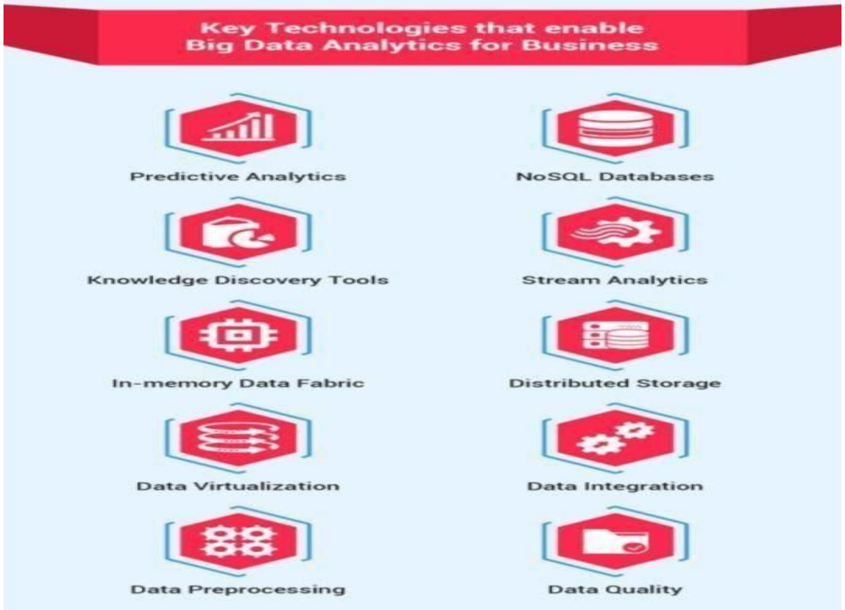
10 Key Technologies that enable Big Data Analytics for businesses

The big data analytics technology is a combination of several techniques and processing methods. What makes them effective is their collective use by enterprises to obtain relevant results for strategic management and implementation.

In spite of the investment enthusiasm, and ambition to leverage the power of data to transform the enterprise, results vary in terms of success. Organizations still struggle to forge what would be consider a "data-driven" culture. Of the executives who report starting such a project, only 40.2% report having success. Big transformations take time, and while the vast majority of firms aspire to being "data-driven", a much smaller percentage have realized this ambition. Cultural transformations seldom occur overnight.

At this point in the evolution of big data, the challenges for most companies are not related to technology. The biggest impediments to adoption relate to cultural challenges: organizational alignment, resistance or lack of understanding, and change management.

Here are some key technologies that enable Big Data for Businesses:



1) Predictive Analytics

One of the prime tools for businesses to avoid risks in decision making, predictive analytics can help businesses. Predictive analytics hardware and software solutions can be utilized for discovery, evaluation and deployment of predictive scenarios by processing big data. Such data can help companies to be prepared for what is to come and help solve problems by analyzing and understanding them.

2) NoSQL Databases

These databases are utilized for reliable and efficient data management across a scalable number of storage nodes. NoSQL databases store data as relational database tables, JSON docs or key-value pairings.

3) Knowledge Discovery Tools

These are tools that allow businesses to mine big data (structured and unstructured) which is stored on multiple sources. These sources can be different file systems, APIs, DBMS or similar platforms. With search and knowledge discovery tools, businesses can isolate and utilize the information to their benefit.

4) Stream Analytics

Sometimes the data an organization needs to process can be stored on multiple platforms and in multiple formats. Stream analytics software is highly useful for filtering, aggregation, and analysis of such big data. Stream analytics also allows connection to external data sources and their integration into the application flow.

5) In-memory Data Fabric

This technology helps in distribution of large quantities of data across system resources such as Dynamic RAM, Flash Storage or Solid State Storage Drives. Which in turn enables low latency access and processing of big data on the connected nodes.

6) Distributed Storage

A way to counter independent node failures and loss or corruption of big data sources, distributed file stores contain replicated data. Sometimes the data is also replicated for low latency quick access on large computer networks. These are generally non-relationaldatabases.

7) Data Virtualization

It enables applications to retrieve data without implementing technical restrictions such as data formats, the physical location of data, etc. Used by Apache Hadoop and other distributed data stores for real-time or near real-time access to data stored on various platforms, data virtualization is one of the most used big data technologies.

8) Data Integration

A key operational challenge for most organizations handling big data is to process terabytes (or petabytes) of data in a way that can be useful for customer deliverables. Data integration tools allow businesses to streamline data across a number of big data solutions such as Amazon EMR, Apache Hive, Apache Pig, Apache Spark, Hadoop, MapReduce, MongoDB and Couchbase.

9) Data Preprocessing

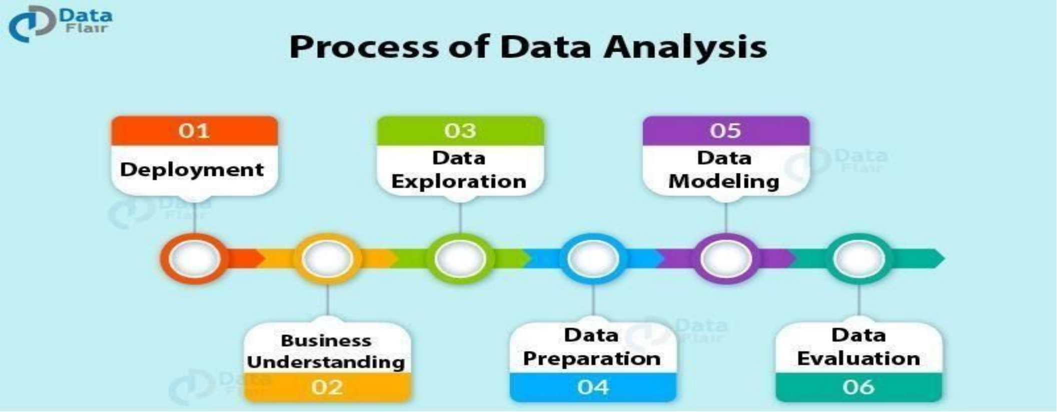
These software solutions are used for manipulation of data into a format that is consistent and can be used for further analysis. The data preparation tools accelerate the data sharing process by formatting and cleansing unstructured data sets. A limitation of data preprocessing is that all its tasks cannot be automated and require human oversight, which can be tedious and time-consuming.

10) Data Quality

An important parameter for big data processing is the data quality. The data quality software can conduct cleansing and enrichment of large data sets by utilising parallel processing. These softwares are widely used for getting consistent and reliable outputs from big data processing.

There's no doubt that Big Data will continue to play an important role in many different industries around the world. It can definitely do wonders for a business organization. In order to reap more benefits, it's important to train your employees about Big Data management. With proper management of Big Data, your business will be more productive and efficient.

Process of Data Analysis



1. Business Understanding

Whenever any requirement occurs, firstly we need to determine the business objective, assess the situation, determine data mining goals and then produce the project plan as per the requirement. Business objectives are defined in this phase.

2. Data Exploration

For the further process, we need to gather initial data, describe and explore data and lastly verify data quality to ensure it contains the data we require. Data collected from the various sources is described in terms of its application and the need for the project in this phase. This is also known as data exploration. This is necessary to verify the quality of data collected.

3. Data Preparation

From the data collected in the last step, we need to select data as per the need, clean it, construct it to get useful information and then integrate it all. Finally, we need to format the data to get the appropriate data. Data is selected, cleaned, and integrated into the format finalized for the analysis in this phase.

4. Data Modeling

After gathering the data, we perform data modeling on it. For this, we need to select a modeling technique, generate test design, build a model and assess the model built. The data model is build to analyze relationships between various selected objects in the data. Test cases are built for assessing the model and model is tested and implemented on the data in this phase.

5. Data Evaluation

Here, we evaluate the results from the last step, review the scope of error, and determine the next steps to perform. We evaluate the results of the test cases and review the scope of errors in this phase.

6. Deployment

We need to plan the deployment, monitoring and maintenance and produce a final report and review the project. In this phase, we deploy the results of the analysis. This is also known as reviewing the project.

The complete process is known as business analytics process.

Crowd Sourcing Analytics

Crowdsourcing data collection consists in building data sets with the help of a large group of people. There are a source and data suppliers who are willing to enrich the data with relevant, missing, or new information.

This method originates from the scientific world. One of the first ever case of crowdsourcing is the Oxford English Dictionary. The project aimed to list all the words that enjoy any recognized lifespan in the standard English language with their definition and explanation of usage. That was a gigantic task. So the dictionary creators invited the crowd to help them on a voluntary basis.

5 Benefits of Crowdsourcing Big Data

Tapping into stores of data to fuel innovation isn't a new concept. Long before we could access unlimited streams of information with a few keystrokes, smart business leaders relied upon a variety of resources to make the decisions that drove their success. The question isn't whether you should be mining big data to improve distributed processing and analysis. It's how to do it most effectively.

Anyone can access the data that's out there. Your ultimate goal, however, is to proficiently apply that data – to draw out the most value quickly and cost-effectively, rather than relying upon a representative sampling that you *hope* hits the mark. With so much to analyze – and so much at stake – crowdsourcing is an optimal solution. When you pair crowdsourcing with the inherent improvements in big data analysis, you get big benefits:

- You capitalize on the human element.
- You save time.
- You save internal resources.
- You take advantage of scale.
- You get real-time analytics.

1. Crowdsourcing big data capitalizes on the human element.

The human touch lets you drill down to insights that empower you to thrive in a customer- driven, demand market. For sentiment analysis and content moderation especially, crowdsourcing wins over machines and software every time. As smart as computers are in some regards, they lack one vital element when it comes to big data: subjectivity. Analyzing data or moderating content from social networks, customer feedback, or comments and reviews with crowdsourced labor produces insightful, accurate and actionable results from real people – not machines.

2. Crowdsourcing big data saves time.

While you're spending your time moving from ideation to fruition when it comes to SEO contentfocus, search relevance, or cross-touchpoint streamlining with product matching and categorization, your competitors are winning over consumers and super-charging *their* sales. The distributed nature of crowdsourcing ensures that your data gets processed at unprecedented speeds – thousands of workers tackle your projects simultaneously to deliver results faster than you could ever achieve in-house.

3. Crowdsourcing big data helps you save internal resources.

You're good at what you do. Your employees are good at what they do. Why *waste* internal resources assigning over-qualified staff big data processes that a crowdsourced workforce can tackle faster and more efficiently? Likewise, on-boarding new staff carries overhead and incidental employment costs that crowdsourcing eliminates. You not only save internal resources, you maximize cost-effectiveness with crowdsourcing.

4. Crowdsourcing big data allows you to keep up with demand as you grow.

No aspect of your business is static. As you grow, your accumulated data grows with you. Without the right tools, tapping into that data can become an insurmountable task. With crowdsourcing, your tasks scale with you. When you have thousands of hyper-specialized crowdworkers available, it doesn't matter how quickly or dramatically your needs increase. Your crowdsourced work force can handle as much as you put on its plate – improving site search functionality, personalizing marketing campaigns, catalog cleansing, moderating content – with consistent quality, speed and insight.

5. Crowdsourcing big data provides you real-time analytics.

In today's competitive online environment, *right now* matters. You can't place consumer demands on hold while you sift through out-of-date data. Using old data is like showing up to the party after everyone's already left. When you use crowdsourcing, sentiment analysis, content moderation, categorization and other data tasks happen in real-time, so you never have to worry about being unfashionably late.

When conducting business online, you must juggle multiple tasks, including creating initiatives that satisfy customers, delivering solid ROI for marketing costs, and consistently generating higher conversion and retention rates for online sales. The prevalent attitude is that customers come first, and there's no shortage of retailers, marketers, websites and mobile apps that confirm that message and deliver on it. You can easily be one of them by utilizing crowd-powered big data insights to drive effective marketing and online sales.

10 Top Big Data Companies in India

Here are those 10 best big data analytics companies in India to work for and to get services. If you are looking for the services related to big data, you may look for these top big data spark companies.

Fractal Analytics

Fractal Analytics is one of the best big data companies in India serving the fortune 500 clients in the domains like CPG, Financial Services, Insurance, Retail, Technology, Life Science, Healthcare, Telecommunication and Media industries. It was founded in 2000 having headquartered in Mumbai, the financial capital of India.

As for now, the company is majorly focusing on Big Data Analytics, AI, and Machine Learning. The company was founded by Srikanth Velamakanni and Pranay Agrawal and very soon become one of the leading analytics company of India.

Total Workforce: 1k-5k
Total Valuation: Around USD 300 million
Number of acquisitions: 4

Impetus Technologies

Impetus Technologies is second in our list which was founded by Praveen Kankariya in 1991. Since then Impetus has expanded its portfolio with the companies like ClearTrail Technologies, Intellicus Technologies, and the products like Kyvos Insights and others.

Impetus majorly deals in the big data services and serve the majority of the top-tier companies. They are considered to be the leading companies in the big data world. Their product Kyvos enables you to deal with the OLAP on top of Hadoop which is a similar tool as Apache Kylin.

Although the company has it's headquartered in Los Gatos, CA, USA but majorly operates from Noida (UP, India), Bangalore (KA, India), and Indore (MP, India).

Total workforce: Around 2k
Total Revenue: USD 60 Million (as per data from 2014)
Number of acquisitions: Having own created Subsidiaries like ClearTrail Technologies Inc, qLabs – impetus, Impetus Infotech (India) Private Limited, iLabs, ProXel, pLabs – impetus, mLabs – impetus.

Mu Sigma

Mu Sigma started as a startup and now is one of the most trusted names in the big data analytics, AI, and machine learning space. Mu Sigma is the world's leading provider of analytics and decision science solutions

Founded in 2004 by Dhiraj Rajaram, Mu Sigma is a top choice of techies from the premier institute in India. This Bangalore-based analytics giant majorly deals with the management consulting to the top organizations across the world. Mu Sigma has received a total funding of more than $210 million in multiple rounds of funding and has a unicorn status.

Total Workforce: Around 3.5k

Total Valuations: USD 1.5 Billion (2017)

Total Acquisitions: NA

Absolut data

Absolutdata is another leading big data company in India offering wide ranges of advanced analytics solution including big data, machine learning, artificial intelligence etc.

Founded in 2001, the company headquartered in Alameda, California, United States. It was founded by Anil Kaul and currently serves globally. The company raised funding in 2012 with the total sum of $20 Million.

Total Workforce: Close to 1000

Total Valuations: Around USD 1.5 Billion

Number of acquisitions: NA

Tiger Analytics

Tiger Analytics is another big data analytics organization in India having a global presence. The company deals in data analytics and predictive modeling. The company was founded in 2011 and is headquartered in Santa Carla, CA, USA.

Tiger Analytics is currently offering their data solutions to multiple fortune 100 clients in the niches like retail, social media, and online advertising sectors.

Total Workforce: 100-200
Total Revenue: USD 10-25 Million per year
Number of acquisitions: NA

Bridgei2i

Bridgei2i is the winner of fast technology 50 India winner in consecutive 3 years (2015-2017) and also an interact in the Gartner magic quadrant. The company was founded in 2011 and this Bangalore based analytics consulting firm majorly deals in sales, retail, customer, and marketing analytics.

It was founded by Prithvijit Roy and has grown multiple folds since then. Bridgei2i has also raised series A round of funding of some undisclosed amount.

Total Workforce: 200-300
Total Revenue: USD 8-10 Million per year
Number of acquisitions: NA

LatentView

The LatentView is another top big data analytics company offering digital analytics, market analytics, and web analytics solutions to the clients. It offers the solutions like data engineering, supply chain analytics, business analytics and more. The company has also won Deloitte technology fast 50 since 2009.

Founded in 2006 by Gopi Koteeswaran, the company is headquartered in Chennai, India. LatentView currently serves the clients in the field of advanced analytics like data engineering, data science, machine learning etc. including the clients like Microsoft, Expedia, PayPal etc.

Total Workforce: 200-500
Total Revenue: USD 2 Million per year
Number of acquisitions: NA

Crayon Data

Crayon Data is another leading analytics company working in the fields of advanced analytics. Started by Vikram Rao in 2012, the company is headquartered in Singapore. Company majorly provides a personalized experience to the customers with the help of data.

So far, Crayon Data has received a funding of USD 5.3 million and the interesting thing is, Crayon Data generates $200k in revenue for each of their workforces.

Total Workforce: 150-200
Total Revenue: USD 35 Million per year
Number of acquisitions: NA

Indix

Indix is majorly an artificial intelligence platform for product data information. The company works with eTailer, Ad Tech, Marketplaces, affiliates, research & analytics, Asset Managers etc. Indix was founded by Sanjay Parthasarathy in 2010 and is headquartered in Seattle, Washington. As per the data available, Indix has received around $35.9 million in various rounds of funding. For every employee at Indix, the company generates around $15.9K.

Total Workforce: 50-100
Total Revenue: USD 1.2 Million per year
Number of acquisitions: NA

Datamatics

Datamatics is one of the top big data companies in India. The company is now a big brand when it comes to big data and serves several fortune companies.

The company majorly deals with data related technologies like Big data, data analytics, Artificial intelligence, machine learning and others. The company is a stock market listed company and has multiple subsidiaries as well as Cignex Datamatics.

Total Workforce: 8000

Total Revenue: USD 44 Million per year

Number of subsidiaries: Cybercom Datamatics Information Solutions Limited, Lumina Datamatics Limited, CIGNEX Datamatics Technologies Limited, Datamatics Vista Info Systems Limited, LDR eRetail Limited, LD Publishing & eRetail Limited, Datamatics Robotics Software Inc.

PromptCloud

PromptCloud Technologies Pvt. Ltd. is a leading web data crawling & extraction company, serving customers across the globe with alternative data to suit their business needs. Based on the Data-as-a-Service (DaaS) model, PromptCloud uses cloud computing and machine learning techniques to offer big data solutions to enterprises.

PromptCloud web crawling service helps businesses get the data they want, the way they need it. It make use of advanced web crawling, web scraping and data extraction techniques to deliver clean and ready-to-use data which powers various business intelligence applications.

As a hosted solution, PromptCloud is ideal when you want data from specified websites. Data can be in the form of reviews, blogs, product catalogs, social sites, travel data, and even real-time tweets from Twitter. Web data extracted is delivered in desired format via a REST-based API. Value-added services include Hosted Indexing, and Live Crawls. Apart from these PromptCloud offer two more solutions, viz., JobsPikr (job data feed provider) and DataStock (historical web data set provider).

As a data solutions company, PromptCloud has clients from all over the world. Vertically agnostic, clients include top names in Fortune 500 companies, start-ups & SMEs from various sectors like E-commerce & Retail, Travel & Hospitality, Finance, Healthcare, Marketing & Business Research, Analytics etc.

Total Workforce: 30
Total Revenue: 1M USD
Number of acquisitions: NA
Founder & CEO: Prashant Kumar

The top big data companies in India list can’t be completed with only these 10 companies and we have hundreds of best big data analytics companies in India and globally

Top Data Science Companies in India

A day before yesterday when I was looking at the financial daily, a news article about the rapidly growing Big Data and data analytics market caught my attention. It reminded me of Gartner’s earlier prediction of Indian business intelligence and analytics software market reaching $245 million in 2017, a whopping 24.4% increase over 2016 revenue of US$206 million. A more detailed analysis of The Hindu BusinessLine forecasts the addition of around 1.80 lakh to 2 lakh new jobs in 2018. Top Data Science companies in India are looking for fresh qualified data analytics resources. Many of the global players in the data analytics have set up development centers in India.

Many of the top Data Science companies in India are ready to raise the bars for salary and compensation for talented data analytics professionals. Most of these hot employers, who are also some of the best Data Science companies in India are building a strong resource pool of data analysts.

Here, I have compiled a list of top Data Science companies in India, which have already proved their presence in the market.

blueOcean
market intelligence

1. BLUEOCEAN:

Ranked among the top Data Science companies in India, Blueocean Market Intelligence helps organizations realize a 360-degree view of their customers through data integration and a multi-disciplinary approach that enables sound, data-driven business decisions. Founded in 2000, it is one of the top Data Science companies to work for.

Through their 360 Discovery approach, Blueocean ensures the comprehensive use of all available structured and unstructured data sources. The qualified resource pool is a delightful mix of analytics, domain expertise, engineering and visualization skills brought together in harmony. BlueOcean has established a rich customer base of leading global organizations in Europe, Asia, Middle-East, and North America.

Already ranked among the top Data Science companies in India, BlueOcean has also been recognized for its remarkable impression in Big Data analytics. In 2017, it was ranked among India's top 10 leading analytics providers by Analytics India Magazine. Minerva (one of the premium products of BlueOcean) has been recognized as the "Emerging Analytics Product Startup of the year" at CYPHER 2017 and Digital Application of the Year award at the CMO Asia Social Media & Digital Marketing Excellence Awards, Singapore in 2016.

2. MANTHAN:

Manthan, one of the top Data Science companies in India, is a pioneer in analytical applications for consumer-facing businesses. The specialty areas include Retail Analytics, Customer Analytics, Consumer Goods Analytics, Advanced Analytics Solutions, Big Data Analytics, Cloud Analytics, Retailer Supplier Collaboration, eCommerce Analytics, Predictive Analytics, and Analytics- driven and location-based marketing

Manthan, one of the top science companies is widely known for its unique products. Manthan engineered Maya, the world's first AI-powered conversational interface for business analytics. With Maya, decision-makers can interact with their data in natural language and perform complex business inquiries through a voice interface. Manthan's other products, powered by AI, cloud, and perspective capabilities are unique in their ability to use machine intelligence to process decision contexts and respond automatically with actions and recommendations to manage every aspect of a consumer business. Headquartered in Bangalore with offices in Santa Clara, London, Dubai, Mexico City, Singapore, and Manila, Manthan is one of the fastest growing data scientist companies in India.

One of the top Data Science companies in India, Manthan is also one of the most awarded analytics innovators among analysts and customers. With over 170 customers across 22 countries, Manthan is among the top data scientist companies in India.

3. DATALICIOUS:

Datalicious, one of the top Data Science companies in India, helps marketers improve customer journeys through the implementation of smart data-driven marketing strategies. Datalicious has a dedicated team of marketing data specialists, offering a wide range of skills suitable for any challenge and cover everything from web analytics to data engineering, Data Science, and software development.

Datalicious, one of the best Data Science companies to work for, offers a wide range of products and services, such as the OptimaHub, SuperTag and Google 360 Suite. The company has its client base mostly in APAC, Southeast Asia, India, Europe & North America.

Ranked among the largest premium reseller of Google Analytics in APAC region, Datalicious works with the whole spectrum of clients from Fortune 500 to early-stage startups. It designs tailor- made Data Science solutions for businesses at any data maturity stage; whether you are a startup looking for help with your overall data strategy or a seasoned player looking for specialists in niche areas like marketing attribution and predictive modeling.

With a 50-member organization that includes an eclectic a mix of analysts, data scientists, and client success managers, Datalicious is one of the fastest growing data scientist companies in India. The company has its head office in Sydney, Australia and much of the major delivery and development happens from Bangalore office in India.

4. FRACTAL ANALYTICS:

Fractal Analytics, one of the top Data Science companies in India, was founded in 2000. The company partners many of the Fortune 500 companies. Fractal Analytics helps top global brands power every human decision in the enterprise by bringing analytics & AI to the decision-making process. Fractal Analytics works with organizations to build breakthrough analytics solutions, set up analytical centers of excellence and institutionalizes data-driven decisions.

Concordia, a Fractal product, enables an organization to be analytics ready. The unique Data Science solution helps organizations in making intelligent and impactful business decisions by harmonizing data from disparate sources using Fractal Analytics’ proven accelerator platform.

Fractal Analytic, one best Data Science companies in India, has a presence across 12 global locations including the United States, UK, and India. Most of its core clients are based in San Francisco Bay area, Greater New York area, London, Mumbai, New Delhi, Singapore, and Dubai.

5. CARTESIAN CONSULTING:

Cartesian Consulting, founded in 2009, in Mumbai, is one of the best-known data scientist companies in India. The global analytics services company helps organizations improve revenues and margins by helping them better utilize their data for business decisions and marketing interventions.

Cartesian Consulting’s work has helped many global brands in retail, QSR, financial services, telecom, eCommerce, and hospitality improve their top lines by 5%-10%, improve margins by 6%-12%, and improve their ability to take better data-driven decisions. Cartesian, one of the top Data Science companies in India, is a fast-growing organization working for over 50 brands in 10 countries. They offer solutions in customer analytics, digital analytics, demand forecasting, recommendation engines, NLP, and text mining.

Ranked among the top Data Science companies in India, Cartesian Consulting has offices in several locations including Mumbai, Bangalore, Gurgaon, Singapore, and San Francisco. It has its client base spreads across APAC and Middle Eastern markets.

6. CRAYON DATA:

Crayon Data, one of the best Data Science companies in India, was founded in 2011. The company enables enterprises to provide ultra-personalized choices to their customers, using their proprietary platform- SimplerChoices. SimplerChoices currently has data that covers 1 billion tastes, 25 billion taste connections, products, 25 million products.

CrayonData, one of the top Data Science companies in India, has offices in Chennai, Singapore, UK Dubai, and the US. Crayon Data was ranked among the top 5 in IBM Watson Mobile Developer Challenge 2014 and a Winner at TiE Silicon Valley's TiE50 Awards in the software vertical. The company, one of the best-known data scientist companies in India, was also selected as one of the top 50 emerging tech companies in India at InTech50 2015.

7. LATENTVIEW:

Founded in 2006, LatentView is one of the top Data Science companies in India. The digital analytics company provides a 360-degree view of the digital consumer, enabling companies to predict new revenue streams, anticipate product trends and optimize investment decisions. Cisco, PayPal, Microsoft, and IBM are among the valued clients of LatentView.

LatentView, one of the best-known data scientist companies in India, has won several awards recognizing its growth and expertise. Very recently, LatentView was named 'Analytics Company of the Year – 2015' by Frost & Sullivan. The company also topped charts for the exclusive list of Advanced Consulting Partners to Amazon Web Services (AWS), for their remarkable expertise and experience in Big Data Analytics. LatentView, one of the top Data Science companies known globally, has also been named a 'cool vendor' in data analytics by Gartner. LatentView is headquartered in Princeton, New Jersey. It has offices in Chennai, Singapore, London, and the US.

8. BRIDGEI2I:

BRIDGEi2i, another known name in Data Science market in India, is a global Analytics Solutions firm and a trusted partner for enabling data-driven business transformation. One of the top Data Science companies in India, BRIDGEi2i enables businesses to contextualize data, generate actionable insights from complex business problems, and make data-driven decisions across pan- enterprise processes to create sustainable business impact. BRIDGEi2i, a top choice for data scientists in India and abroad, strives to generate value to the clients in each stage of their analytics journey from information to insight to impact.

BRIDGEi2i, fast growing as one of the top Data Science companies in India, has a rich client base in several industries, including banking and financial services, insurance, technology, retail, and consumer packaged goods. ExTrack (which is a customer experience management platform), S- Reco (which is a sales recommendation engine) and M2 (a model governance and monitoring solution), are some of the key analytics platforms of BRIDGEi2i.

BRIDGEi2i, one of the best Data Science companies in India, continues to grow with an employee strength of over 270. Headquartered in Bangalore, India, the company also has offices in Fremont, Dallas, Chicago, Seattle, and Boston.

9. HANSA CEQUITY:

Hansa Cequity, one of the fastest growing data scientist companies in India, helps companies build intelligent, intuitive, and real-time customer relationships. The company does this by leveraging the power of technology through proprietary and best-in-class marketing automation and analytics platforms.

Hansa Cequity’s ability lies in bringing in multidisciplinary teams of specialists with rich expertise in product/services marketing organizations, advertising and direct marketing agencies, analytics companies, technology consultancies, contact centers and digital and creative agencies enables them to stand out in the market by offering unique value propositions for their clientele.

One of the top Data Science companies in India, Hansa Cequity, is known for its experiences in customer strategy, data management, analytics, digital campaign management and social media strategies. From simplistic data exploration to building complex analytical models using advanced machine learning/ AI algorithms Hansa Cequity helps organizations identify the right engagement strategies for their customers. It offers services in consulting services (customer strategy, marketing, data management, analytics, digital campaigns management and social media), data management platform, customer analytics and insights platform, and marketing optimization platform. Headquartered in Mumbai, India, Hansa Cequity has plans of expanding to overseas territories soon.

TEG Analytics
INSIGHTS @SPEED OF BUSINESS

10. TEG ANALYTICS:

TEG Analytics, though a newer player in the fray, is fast catching up. Growing rapidly, as one of the best data scientist companies in India, TEG Analytics is a data-science-as-a-services company, helping organizations make decisions at the intersection of business, technology, and applied mathematics.

TEG Analytics, of the best-known names for Data Science, focuses on the alignment between the speed of business and speed of insights. Its proprietary FutureWorks, built on open source/ Big Data platform, provides hassle-free last mile analysis, reduces the time to market and increases the adoption of analytics throughout the client organization. It is closely associated with well-known brands in industry verticals like retail, CPG, healthcare, BFSI in shaping strategy powered by analytics.

TEG, one of the best top Data Science companies in India, has several premium products to its credit, including HeathWorksTM. HeathWorksTM is a solution powered by Tableau and publicly available healthcare data, helps payors develop an ability to predict in-market performance and reads the entire industry landscape to help them understand the competitiveness of their products. DigiWorks, another product, TEG Analytics consists of a suite of analytics models to help companies with strategic, operational, and tactical decisions in their digital efforts.

Data Science companies in India are growing in numbers and many of the startups are doing better than software giants in the market for decades. That is because Data Science offers lucrative career options. so, are you game? You too can be part of the best Data Science companies in India. You may start as a data analyst or become a data scientist earning some more years of experience. Later you may move on to advisory roles in government or private sectors. Many well-known Data Science influencers in India and abroad have started as a data scientist and held leading positions in top Data Science companies in India.