Block

3

BUSINESS ANALYTICS

Editorial Team

Prof. R. Prasad
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sanjay Fuloria
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sindhuja
IFHE (Deemed-to-be-University), Hyderabad

Dr. Nasina Jigeesh
IFHE (Deemed-to-be-University), Hyderabad

Content Development Team

Prof. Anirudh Prakhya
IFHE (Deemed-to-be-University), Hyderabad

Dr. Y. V. Subrahmanyam
IFHE(Deemed-to-be-University), Hyderabad

Prof. R. Muthukumar
IFHE (Deemed-to-be-University), Hyderabad

Prof. V. Srinivasa Murthy
IFHE (Deemed-to-be-University), Hyderabad

Dr. K. Veena
IFHE (Deemed-to-be-University), Hyderabad

Prof. KV Murali Mohan
IFHE (Deemed-to-be-University), Hyderabad

Prof. M. Venkata Dharma Kumar
IFHE (Deemed-to-be-University), Hyderabad

Proofreading, Language Editing and Layout Team

Prof. Jayashree Murthy
IFHE (Deemed-to-be-University), Hyderabad

Mr. K. Venkateswarlu
IFHE (Deemed-to-be-University), Hyderabad

Ms. C. Sridevi
IFHE (Deemed-to-be-University), Hyderabad

Ref. No. BDCA-SLM-IFHE – 122022B3

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

Our E-mail id: cwfeedback@icfaiuniversity.in

Centre for Distance and Online Education (CDOE)
The ICFAI Foundation for Higher Education
(Deemed-to-be-University Under Section 3 of UGC Act, 1956)
Donthanapally, Shankarapalli Road, Hyderabad- 501203

BLOCK 3: BUSINESS ANALYTICS

Data analytics is the science of analyzing raw data in order to arrive at useful conclusions. The techniques and processes of data analytics have been automated with the help of mechanical processes and algorithms. These algorithms work on raw data to make it suitable for human consumption. Data analytics helps optimizing business performance. Analytical tools like statistical approaches and data mining help to reveal many useful business facts out of the collected big data.

Unit 10 – Collecting big data has high relevance when multiple analytical decisions are to be arrived from the data. For instance, marketing data needs analytics to arrive at many vital decisions. The unit on *Analytics in Database Marketing* deals with what analytics is, its connection to database marketing, usage, data mining, issues in analytics, four pillars of analytic competition, CRISP-DM model, analytics as a competition edge, some typical tasks performed by analytics for the benefit of the organization.

Unit 11 – As data is collected, there are chances that some noises (meaningless information) too creep in. Noises need to be identified and segregated to make the data useful. *Business Analytics Techniques* covers the range of data analysis, geospatial intelligence, details on noise, processes of analytics, free creation to utilization of business analytics.

Unit 12 – From basic analytics operations, one shall move to the advanced technology available to analyse the big data. The unit on *Data Visualization and Modelling* spans visualizing the data analytics, data visualization as organizational way for actions, from sampling to using the data, thinking out of the box, 360^0 modeling, data processing speed, availability of technology, beyond the tools to analytics applications. At this stage learner is comfortably placed at understanding and using both big data and related analytics for business decision making.

Unit 10

Analytics in Database Marketing

Structure

"Data analytics is the future, and the future is NOW! Every mouse click, keyboard button press, swipe or tap is used to shape business decisions. Everything is about data these days. Data is information, and information is power."

- Radi, data analyst at Centogene

10.1 Introduction

When organizations are able to collect voluminous primary data from various sources, their business decisions can be more analytics based leading to higher probability of operational success.

In the previous unit we discussed about information management covering, foundation of big data, computing platforms that can handle big data & analytics, big data computation, storage and computational limitations. We have also discussed about the emerging technologies related to big data. Having collected and stored data for various requirements, we now need to analyze the large data.

In this unit we will learn about what is analytics, its competition edge and some typical tasks performed by analytics. We will also learn about database marketing and data mining.

10.2 Objectives

After going through this unit, you will be able to:

- Explain analytics and big data
- Relate to database marketing
- Recognize data mining
- Discuss the CRISP-DM model
- Describe the four pillars of analytical competition
- Illustrate analytics as a competitive edge
- Identify tasks performed by analytics

10.3 What is Analytics?

According to Davenport and Harris analytics is "The extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and add value."

When a business is said to have analytics, it means it uses models based on data to make business decisions, instead of making decisions based on 'gut feel' or 'instinct'.

Along with analytics the terms database marketing, big data and data mining are often heard. It would therefore be pertinent to understand these three concepts as they form adjacent concepts to analytics.

Example: Use of Analytics at a Global Specialty Toy Retailer

A global specialty toy retailer realized that, they were losing out considerable revenue through their mobile website during customer orders through carts. They had to assess the contributing factors. More so, as they were approaching holiday season, which gave more than 40% of yearly sales. To provide analysis, recommendations, implementing the solution within a short amount of time, they partnered with Blast Analytics & Marketing. Blast audited their marketing channels, covered organic search, allied marketing links, company product reviews, the various social media under use, and paid search and arrived at the problem to be website related. By performing 'funnel analysis', Blast identified more macro-level issues which led to an increased abandoned carts. Blast used 'Adobe Analytics' error tracking, identified the page users were on, and the type of error message they received.

Contd....

They 'Implemented field tracking', mined to more granular look and located the exact field where users abandoned the process. These analytics benefitted the toy retailer.

Source: https://www.blastanalytics.com/adobe-analytics-case-study (case study) Accessed on 26/08/22

10.4 Database Marketing

Peppers and Rogers defined database marketing as - "It is the process of building, maintaining and using customer database and other databases for the purpose of contacting and transacting".

They go on to define Customer Database "As an organized collection of comprehensive data about individual customers or prospects that is current, accessible and actionable for such marketing purposes as lead generation, sale of product or service, or maintenance of customer relationships."

Without a customer database there is little one can do in analytics. That is the reason why capture of transaction data is paramount. That is also a reason why most CEM advocates state that having an ERP to capture the data is extremely desirable.

Example: Subaru Automobiles Connects to Data for Understanding the Customer Journey

Subaru was the twenty-first largest automaker by production worldwide in 2017 and was the automobile manufacturing division of Japanese transportation conglomerate Subaru Corporation. Subaru needed an integrated analytics environment, consolidating data 'campaign orchestration, web analytics, data management platform (DMP), and internal CRM systems'. Merkle partnered with Subaru and Merkle used Data Accelerator platform and analysed Subaru's data, like 'CRM data, email and direct mail, analytics click stream data, and DMP digital log data'. Merkle also provided for a cloud-based big data environment, capable of storing, loading, and exposing all the data elements needed for analysis. The feed inventory was established, and Merkle and Subaru's stakeholders jointly worked for smooth deployment leading to use of database for customer understanding benefitting Subaru.

Source: https://www.merkle.com/thought-leadership/case-studies/how-subaru-quickly-connected-its-data-understand-customer-journey (case study) Accessed on 26/08/22

10.5 Database Marketing Usage

Typically database marketing has been used for –

i. Prospecting and screening

Prospecting involves identifying a group of people that are likely to be interested in the product. Typically prospecting is generated through development of Target segments. Prospects have to be screened for credit ratings in case of some of the products like cards and loans.

ii. **Cross Selling**

Cross selling refers to the marketing program which desires to induce existing customers to buy different products. For example, a firm has many customers for its grocery products. It has developed a new line of cosmetics and wants to market the same to its existing customers.

iii. **Up Selling**

In up selling the firm wants its existing customers to use a premium or value added product. For example, a hotel has built up sizable loyal customers since its inception. The hotel has now built a new annex that features premium rooms with personalized service that is designed to bring in more revenue. It wants to market these annex rooms to its existing loyal customers.

iv. **Market Basket Analysis**

Market basket analysis looks at the shopping basket of its customers to detect patterns. Are there products which appear to be purchased together? Are there products which are never in the same basket? Which sections of the retail store is frequented? Based on the analysis the firm can design marketing actions. These could be adjacent placement of products, product bundling, coupon plans etc.

v. **Attrition and Churn**

Are there customers with high probability of switching to other suppliers or service providers? Are there customers who would change their purchase levels (downwards)? If a firm could identify such customers from an analysis they can create tailor made programs to prevent them. For this to happen, these customers need to be identified and convinced to stay.

vi. **Fraud Detection**

Typically, fraud detection involves developing patterns and identifying potential frauds and problems. These can be then monitored closely to minimize fraud without change in service levels.

Example: CRM with Adobe at Family Entertainment Company

A global family entertainment company migrated their CRM activities to corporate level. This tremendous effort required the hiring of an international CRM team, and tools or vendor partnerships to execute on the new strategy. As the databases were updated on a monthly basis, they tended to be outdated, inflexible with significant gaps both in accuracy and efficacy.

Contd….

The current campaign application was out dated and could not support making innovative approaches difficult to handle. Current approach to campaign execution, needed manual intervention, leading to reduced productivity and missed opportunities as well. It also lacked email response data, leading to failed accurate response attribution and measure on audience counts. Having realized these, they implemented best Foundational Marketing Platform (FMP) with best technology supporting their strategic scalability needs. Adobe Campaign was deployed and integrated with the core solution and existing email service provider and Tableau Desktop and Tableau Server was integrated to support advanced campaign reporting and support crucial business intelligence at the family entertainment company.

Source: https://www.merkle.com/thought-leadership/case-studies/best-class-crm-adobe (case study) Accessed on 26/08/22

Activity 10.1

You are the marketing manager of a chain of retail store. Discuss briefly how you would build a dynamic data base for the marketing department of the retail store.

10.6 Types of Big Data

Typically, a firm uses data generated from operations. This is usually termed as (for want of a better name) “small data”. However, with the expansive growth of social media, data volumes have exploded. Data from social media tends to be unstructured. For instance, market intelligence company IDC predicts the world’s data will grow to 175 zettabytes (One zettabyte equals 1 byte followed by 21 zeroes) in the year 2025. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.

Another aspect of big data is that it is unstructured. If data has to be used it should be in a format that makes it amenable for analysis. A format that is usable for analysis is termed structured. Since data is culled from various independent sources, it is in varied formats and hence is not easily amenable for analysis. This is called unstructured data.

Big data therefore is data that is enormous and unstructured.

Check Your Progress - 1

1. What do you call a collection of comprehensive data about individual customers?
   a. Customer database
   b. Lead generation
   c. Sale of product
   d. Sale of service
   e. Maintenance of customer relationships
2. What does the ability of analytics to keep models up to date by considering changed circumstances after developing an initial model make it?
   a. Distinct capability
   b. Hard to duplicate
   c. Provides a unique edge
   d. Can be used in other instances
   e. Renewable
3. What should a firm have to leverage the analytics program?
   a. Distinct capability
   b. Unique
   c. Duplicate product
   d. Duplicate service
   e. Cross boundaries
4. Which goal translates the business objectives in terms that are more technical and research oriented?
   a. Business objectives
   b. Assess situation
   c. Data mining
   d. Project plan
   e. Time
5. In successful analytics based firms, analytics are advocated primarily by which team?
   a. Supply chain management
   b. Customer services
   c. Senior management
   d. Legal team
   e. Administrative

10.7 Data Mining

The most prevalent definition of data mining (Putler and Krider) says it is the "Process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques".

An interesting aspect of the definition is the stated difference between pattern recognition and statistical / mathematical techniques. This is because the definition implies the use of such techniques as machine learning.

Therefore, one can sense that there are at least two categories of data mining techniques. The critical difference is model building (this will be elaborated later). As an introduction we can define the two categories as –

1. Model based on hypothesis: In this category of analysis, the analyst assumes a model based on previously found patterns and logical deductions. The analyst then tests the model using the data. Analysis can determine if the model is valid (statistically significant).
2. Machine learning model: In this category the analyst does not assume a model before an analysis. The techniques used (such as Machine Learning) will extract a model from the data.

Example: Data mining at Cerner Corporation

Cerner Corporation was a leader well known for long time in the healthcare IT space. Their services were utilized in over 14,000 global medical facilities covering 'hospitals, integrated delivery networks, ambulatory offices, and physicians' offices'. The company was moving its focus on to Electronic Medical Records (EMR). Cerner`s aims were: reduce cost, increase delivery of healthcare efficiency, improve patient outcomes. The firm was building a solution on a Big Data platform powered by a Cloudera Enterprise Data Hub (EDH). Cerner needed the ability to iterate fast on the search processing algorithms. Cerner started to leverage SAS on Hadoop for deep data science to build prediction models for reducing hospital readmissions. Cerner had taken full steps to ensure the security and data integrity of its Big Data. It addressed the need for providing a mechanism for threat mitigation, with viable data management technology. The uniqueness about Cerner's EDH was that it integrated data from unlimited sources to build complete picture of a given patient's condition or trend. The centralized hub could predict the probability of a discharged patient, returning for re-admission for the same or a similar condition.

Sources: i) https://www.datamation.com/big-data/data-mining-use-cases/ February 13, 2022 Accessed on 26/08/22

ii) https://www.cloudera.com/content/dam/www/marketing/resources/case-studies/cloudera-cerner-casestudy.pdf?daqp=true

10.8 Issues in Analytics

Big data is therefore a name given to large volume of unstructured data. Data Mining is an analytical process typically using data from business operations. Since data mining looks at historical data to predict future behavior, it is not an error free process. However, by following a process one can eliminate characteristic errors that can seep into the analysis process.

The common errors that occur in data mining can be:

1. Arriving at a conclusion (based on data mining analytics) when it is not true.
2. Arriving at a correct but useless conclusion.

Sometimes patterns seem to emerge from data but they are spurious. It is just like you see a lion, or balancing scale or a fish when you look at the stars. The stars seem to form a pattern, but they can be million light years away from each other. These spurious patterns are often the bedrock of conspiracy theories. For example, there was a popular notion that if the batsman X scores a century, then India will lose the cricket match. This notion was eventually debunked by statistics, but it is still believed to be true by many.

A car showroom conducted a survey amongst its customers to find out how they rate the departments of the firm. The customer has interacted with front room sales, credit executives, trade-in executives, service executives and call-center personnel. The survey concluded that the front room sales were doing a fantastic job. Customers rated them very high. Other departments had customer reservations. While this was welcomed by the management, the conclusion flew in the face of the fact that sales were stagnant. A keen observer pointed out that since the survey was done on actual customers, sales personnel cannot have a low rating. Customers would have switched to a different outlet if they were not happy. The survey was then expanded to include potential customers who had considered the showroom but bought a competitor product. In the expanded survey results were more meaningful. Hence one has to be careful in analyzing the data which may not represent the relevant population.

A grocery store conducted a detailed analysis of the basket that consumers purchased from their outlet. The analysis concluded that people who bought bread usually buy butter and/or eggs together. This is a correct conclusion, but it is already a known. Hence this finding is useless.

A cosmetic firm analyzed their loyal customers for patterns. The most significant finding was that age seems to be correlated to qualifications. Again this finding is correct but useless since the conclusion is trivial.

An appliance firm's analysis arrived at the conclusion that couples who buy appliance X are likely to be separated in three years after purchase. This conclusion may be statistically correct but is useless since it cannot be used by the firm in an ethical way. One cannot advertise the fact that if you want to be separated then you should buy this Appliance.

Example: Analytics accelerator at McDonald

McDonald's, largest restaurant chain was focusing on social media activities which help to 'push the sales, multiply the engagement, and build more positive brand sentiment'. They partnered with the Analytics Accelerator at Wharton Customer Analytics and arrived at three objectives: Identify the impact and differences between: "organic social media activity and paid social activity", and maximize sales by optimized organic social media strategies, establish primary drivers which engaged social media optimized encouraging posts, and Identify a metric and drive brand affinity.

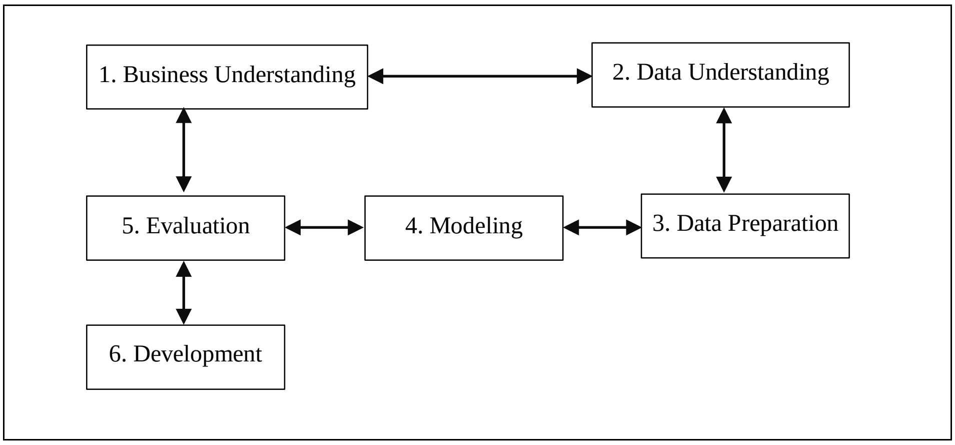
The team identified significant predictors measured through metrics and analyzed length of each media post, type used and product associations, which impacted sales. The Analytics Accelerator team was able to uncover a number of valuable insights regarding McDonald's social media engagement: Posts with videos and carousels (interactive media types) outperformed images and text (stationary media types) , Posts with explicit products like Big Mac had better engagement rates, Cultural or celebrity-related keyword Posts had higher engagement rate.

Source: https://wca.wharton.upenn.edu/research/student-projects/social-media-users-are-lovin-it/ (case study) Accessed on 26/08/22

10.9 CRISP-DM Model

The CRISP-DM model for analysis was developed in order to standardize a process for analytics. The model is hierarchical in nature. It has six phases of analysis and in each phase it has sub-levels (called Generic Tasks). Figure 10.1 presents CRISP-DM phases

Figure 10.1: The CRISP-DM Phases



Source: ICFAI Research Center

The top level of the CRISP-DM model consists of the six phases as outlined in Figure 10.1. The analysis goes through the phases in order but not strictly. Most of the time, the analysis can go back and forth between the phases till clarity emerges. The goal is to get the analysis to be deployed (phase 6).

Phase 1: Business Understanding

CRISP-DM describes this phase as "This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives."

In this phase the analyst understands the business. S/he understands the firm's objectives, sorts out the tangle of objectives in terms or priorities and impact and understands the constraints (limits) of the analysis. This phase has four generic tasks. Each of the generic task is detailed in the following Table. The tasks, descriptions and the desired outputs are detailed in the table as an example. The student can look into the CRISP-DM document (given in the reference section) for detailed generic tasks and outputs of other phases. Table 10.1 describes tasks, description and output phases from CRISP-DM

Table 10.1: Tasks, Description and Output Phases from CRISP-DM

#	Generic Tasks	Description	Outputs
1	Determine Business Objectives	In this task, the analyst understands the business process and formulates the business questions which have to be answered by the analysis. For example, the primary question may be – How to retain customers? The secondary questions may be 1. Identify customer profile of those in danger of dropping out. 2. Will lower interest rates and service charges lead to better retention? (credit cards)	1. Business Background Report 2. Business objectives 3. Success Criteria Example – lower churn rate from 10% to 3%

Contd….

2	Assess Situation	In this task, the analyst looks at all the resources available.	List of resources available including people, data, software, computing resources etc. List of resources required Assumptions Risks & Constraints Glossary Cost Benefit Analysis
3	Determine Data mining goals	The data mining goal translates the business objectives in terms that are more technical and research oriented. For Example – Predict how many customers will surrender their credit card based on their usage, interest terms, service charges, late payment history, qualifications and employment levels.	1. List of Data Mining Goals 2. Success criteria. For example – customer profile will predict 90% of the drop outs.
4	Project Plan	Plan for achieving the data mining objectives including time and costs. The task includes a preliminary analysis identifying the tools and techniques used for the project. This is very essential and will guide the analyst for the next phase.	1. Project Plan 2. Initial Assessment of tools and techniques.

Source: ICFAI Research Center

Phase 2: Data Understanding

In this phase the analyst looks at –

- What data is readily available in the organization, what are the access levels, does the data meet the initial tools and techniques identified in the previous step, identifying (with costs and timelines) third party data that is relevant to the project.
- The data is then explored further to look at the quality and format. Format analysis includes understanding the coding and the mathematical nature of the data (categorical, ordinal, interval or ratio). The total amount of data is also quantified and captured.
- Simple analysis like descriptive statistics (frequency counts, means, histograms, scatter plots) gives a feel for the data and variables in the business.
- Data adequacy is assessed. Is the data enough to reach the stated objectives?

Phase 3: Data Preparation

In this phase the analyst performs the following tasks:

- Cleans the data: There may be missing values, or values which are not correctly entered at the transaction time. For example, the people managing the process would not have queried the customer on his/her anniversary date but would enter 1st January by default. Simple Analysis carried out in the earlier section would throw up such problems. There are no standard solutions to such problems. Sometimes the data may have been captured in some other database.
- Constructs the data: New variables (which are required in the analysis) are constructed from existing data. For example, age may be constructed from date of birth. Grades in qualifications may be constructed from data on qualification.
- Integrates data: Data which are available in different places (hard or soft copies of databases) are brought to one location and integrated to a single database.
- Formats the database: Formatting helps efficient working of the tools.

Phase 4: Modelling

The tasks in this phase are follows-

- Select the technique: Refine ad detail the technique identified in the first phase. For example – Multiple Regression Analysis using log-log scales.
- Test design: Usually the model is not developed on the entire data. A portion of the data (random) is kept aside for testing. The analyst develops the model (build set) on a portion of the data and tests on another portion (test set).

- Generate model(s): For example, a customer who has had three consequent late payments but spends less than Rs 5000 per month is not likely to renew his credit card. Models generally give relationships between various variables and the variable of interest.
- Assess models: If there is more than one model then qualify each with validity and reliability statistics.

Phase 5: Evaluation

The tasks in this phase are as follows-

- To evaluate results from modeling. Has the model produced the desired business and research objective? If the answer is yes, then this model is called approved models.
- To review various phases of the process for flaws and possible problems and validity.
- To determine the next course. If the process is complete, then move to deployment. If further analysis is required, then revert to that stage and iterate.

Phase 6: Deploy the Model

In this phase, the analyst will do the following-

- Describe how to use the model with steps and strategies.
- Describe how to monitor and maintain the model and strategy.
- Prepare a final report and present the same to the client (internal or external).
- Document the entire process.

Example: Crisp DM Methodology at Verizon

Verizon was an American telecommunications company offering wireless products and services. They integrated CRISP-DM model with Tableau and ran applications using data mining. The company's analytics team specifically used the Tableau visualization tool for analyzing sales and could cut customer support calls by 43 percent, resulting in enhanced customer experience.

Phase 1 – Business Understanding: reduced in the number of customer calls by 43 percent for increased customer experience.

Phase 2 – Data Understanding: Extracted and Aligned from various platforms like Oracle, Hadoop, and Terawatt.

Phase 3 – Data Preparation: The team used the 'Tableau Prep Builder' to visually combined data, and structured the data correctly into union sets for Scalability and automation.

Contd….

Phase 4 – Modeling: A decision tree analysis or a neural network generation was used.

Phase 5 – Evaluation: The dashboards on Tableau were checked for detailed view of the customer call scenario and the reasons behind it.

Source: https://www.researchgate.net/publication/344777485_Integrating_Crisp_DM_ Methodology_for_a_Business_Using_Tableau_Visualization July 2020, Accessed on 26/08/22

Activity 10.2

You are the analyst in charge at an e-com selling giant like Amazon. Discuss the steps in implementing CRISP –DM model in your organization.

10.10 Characteristics of Firms with Analytics Approach

Not all firms can reap the benefits of analytics. Research has found that successful firms have four distinct characteristics (Davenport and Harris) -

1. The firm has a distinct capability: The capability is something the firm does better than everybody else and the capability is important to the customers. For example, it can be predicting customer preferences, supply chain management, customer services, choosing the right employees etc. This capability is supported by analytics. It is important to note that a firm without a distinctive capability will not get great benefits from an analytics program. Analytics serves the strategy. It is NOT a strategy by itself.
2. The firm ensures that data available throughout the organization is available for analysis. Often data is captured by various departments of a firm and is kept in silos and not available for others. The firm has to break boundaries.
3. Senior Management Commitment: Successful firms have shown that analytics usage and deployment is due to the commitment by senior management. Usually the CEO is the primary advocate of analytics. Davenport and Harris write about a CEO who “Demonstrates how a CEO—and ideally an entire executive team—who constantly pushes employees to use testing and analysis, and make fact-based decisions, can change an organization’s culture. He’s not just supportive of analytics—he’s passionate on the subject.”
4. Have big ambitions: Often CEOs expect the analytics to have a “Scale and scope of results from such efforts. It should at least be large enough to affect organizational fortunes. Incremental, tactical uses of analytics will yield minor results. Strategic, competitive uses should yield major ones.” (Davenport and Harris).

Example: Data Analytics at Walmart

Walmart **-** The American multinational retail company, had over 11,500 stores in 27 countries globally and e-commerce websites. Walmart had created ‘Data Café’ – a state-of-the-art analytics hub to analyze 2.5 petabytes of data from 1 million customers every hour. Data Café could work with 40 petabytes of transactional data, to model, manipulate, and visualize. Walmart analysed over 100 million keywords to assess customer views on social media, to understand customer behavior, likes and dislikes. Walmart used Python, SAS, and NoSQL Cassandra and Hadoop to derive business insights for improved customer satisfaction. With these data analysis techniques, Walmart could improve management of their supply chain planning and operations, optimize on product assortment, personalize the shopping experience for individual customers, and also share apt product recommendations to their customers.

Source: https://www.simplilearn.com/tutorials/data-analytics-tutorial/what-is-data-analytics, July 2022, Accessed on 26/08/22

10.11 Analytics as a Competition Edge

Analytics can support and provide the competitive edge to a firm. It ensures that the distinct capability of the firm becomes a key differentiator from its competitors.

Analytics is hard to duplicate: You can duplicate a product or a service but a process and culture is difficult to be duplicated.

Analytics provides a unique edge: Every firm has to develop a distinct capability supported by analytics. There is no tried and tested procedure for this. It depends a lot on firm specific variables such as the ones mentioned in the pillars of analytical competition. The firm’s culture and the personnel are drivers and hence Analytics is unique. The uniqueness again makes it hard to imitate.

Analytics can be used in other instances: An analytical program can cross boundaries from CEM to HRM for instance.

Analytics is renewable: Strategies and analytics are not static. They change over times. Hence the capability is just not the result of analytics but the ability to perform them again and again in changing circumstances.

Example: Team Liquid gained Competitive Edge with Data Analytics

Team Liquid was a renowned Netherlands and US-based esports company. They were always on innovation path to remain the world’s premier esports franchise.

Contd….

They were the first esports team to build a state-of-the-art training facility where their athletes balance carefully calibrated practice sessions with healthy meals, along with sports psychologists, personal trainers, and massage therapists. Team Liquid focused on software and data in partnership with SAP Business Technology Platform (SAP BTP) as their next innovation frontier. This had given them a critical edge over the competition.

Team Liquid had hundreds of data points per second for analysis. The data was shifted from spreadsheets into SAP HANA Cloud, and made it readily available for real-time analysis. Thus, with SAP BTP, preparation time for a match was brought to one hour from four days. The tool gave in real-time through predictive analytics, the other team's composition, and terms of their strategy, giving Team Liquid scope to respond and react."

Team Liquid prioritized two of their top-tier games—*Dota 2* and *League of Legends. I*mproving performance by only 1% meant prize money of an extra $50 million USD. Team Liquid had seen the power of data analytics and was expanding to other esports, for better performance of their players.

Source: https://blogs.sap.com/2022/08/15/team-liquid-hones-their-competitive-edge-with-data-analytics/ August 15, 2022 Accessed on 26/08/22

10.12 Some Typical Tasks Performed by Analytics

Some of the successful applications of Analytics (Davenport and Harris):

The ability of the software to recommend movie titles based on past viewing habits is a cutting edge model. The organization makes preference data available for anyone. There is a handsome reward if you can make a model that can beat the organization's model in predicting movie preferences.

A retail shop uses groups to classify customers according to lifestyle. For example, a female shopper who buys on a weekly basis, tries out items which are on sale and uses coupons which are specific to a particular category (Value Conscious). A male shopper visits a restaurant frequently, to eat items. He does not change preferences despite price or promotions. He represents another type (Convenience). Promotions and coupons sent are not the same for these two groups.

David Bell states that, Harrah's (a casino operator) is able to tell "Who is coming into the casino, where they are going once they are inside, how long they sit at different gambling tables and so forth. This allows them to optimize the range and configuration of their gambling games".

A telecom firm discovered through analytics that many customers with unpaid bills were not defaulters but customers who had issues with the bills. The firm changed the way these customers were handled. Instead of sending bill collectors to them they sent service staff to settle the issues and retaining them.

A telecom operator who is specialized in data, pioneers the use of artificial intelligence software to provide subscribers with the content they want in advance. The model uses previous behavior to predict preferences. The preferences are then placed in such a way that the users can have easy access.

Example: Analytics help Retail Company Analyze Customer Behavior

A retail company had to manage big amounts of data coming from various digital channels it operated on. They built on processes to make time-sensitive decisions primarily on ad campaigns, customer promotions and varied other sales aspects to efficiently Sprocess the vast amounts of data, its mobile application and over a million global customers. The company WNS helped them with their proprietary, big data analytics platform BrandttitudeTM, featuring exploratory data analytics, predictive modeling and self-service analytics delivering compelling metrics. It provided accurate and actionable insights through a unified view of disparate data. BrandttitudeTM included: ‘Point of sale, Campaigns, Mobile application’ with identified 150+ Key Performance Indicators (KPIs) for retailers. The retail company could gain more visibility into all siloed data sets. Targeted campaigns were designed by company’s marketing team and determined the product sales volume, customer behavior and allied engagement on its mobile application. BrandttitudeTM helped the company to identify cross-selling products to design effective digital campaigns. The platform’s predictive analytics measured the impact of retail initiatives, covering multiple stores, leading to better forecasting for optimal inventory management. Some other benefits realized by the company were: ‘Analysis and reporting time reduced from few days to few minutes, Easy access of these reports for all decision-makers, Integrated, curated and harmonized data store’.

Sources: i) https://s3.wns.com/S3_5/Documents/CaseStudy/PDFFiles/7066/37/Big-Data-Analytics-Retail_Case-Study.pdf

ii) https://www.wns.com/perspectives/case-studies/casestudydetail/195/big-data-analytics-helps-retail-company-analyze-customer-behavior-and-build-targeted-marketing-campaigns (case study) Accessed on 26/08/22

<u>Check Your Progress - 1</u>

6. What do you call the process of identifying missing values or handling default values?
   a. Cleans the data
   b. Constructs the data
   c. Integrates the data
   d. Formats database
   e. Modelling

7. What are the two sets of data partition?
   a. Multi Regression Analysis
   b. Build, Test
   c. Generate Model
   d. Assess Model
   e. Evaluation
8. Which algorithm has the ability of the software to develop models without specifying the variables of interest?
   a. Model based on patterns
   b. Model based on logical deductions
   c. Machine learning
   d. Statistical technique
   e. Mathematical technique
9. What is Big data?
   a. Small data
   b. Structured
   c. Semi-structured
   d. Enormous and Unstructured
   e. Data mining
10. What do you call analyzing the components of purchase often made together?
    a. Prospecting and screening
    b. Cross Selling
    c. Up Selling
    d. Market Basket Analysis
    e. Attrition and Churn

10.13 Summary

- Analytics provides an analytical underpinning to a firm so that it can leverage its strategic capability to provide a unique, hard to replicate competitive edge.
- We looked at the phases of a structured data analysis project and understood which firms benefit the most by using analytics.
- In addition, data mining concepts, issues in analytics are also covered.
- A model for analytics is added and the four pillars of analytics competition is also discussed.
- Application and simple tasks performed by analytics is discussed for the benefit of the learner.

10.14 Self-Assessment Test

1. Write down your understanding of analytics
2. Express what is database Marketing and usage
3. Explain approaches to data mining from big data storage
4. What is the use of CRISP-DM Model
5. Discuss Four Pillars of Analytic Competition

10.15 Glossary

Analytics (Davenport and Harris): “The extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and add value”.

Customer Attrition: Also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

Data Mining (Putler and Krider): “Process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques”.

Database Marketing (Peppers and Rogers): “It is the process of building, maintaining, and using customer database and other databases for the purpose of contacting and transacting”.

Machine learning: A scientific discipline that explores the construction and study of algorithms that can learn from data.

10.16 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

10.17 Answers to Check Your Progress Questions

1. (a) **Customer Database**

   It is an organized collection of comprehensive data about individual customers or prospects that is current, accessible and actionable for such marketing purposes as lead generation, sale of product or service, or maintenance of customer relationships.

2. (e) **Renewable**

   Strategies and analytics are not static. They change over time. Hence the capability is just not the result of analytics but the ability to perform them again and again in changing circumstances.

3. (a) **Distinct Capability**

   Analytics can support and provide the competitive edge to a firm. It ensures that the distinct capability of the firm becomes a key differentiator from its competitors.

4. (c) **Data Mining**

   The data mining goal translates the business objectives in terms that are more technical and research oriented.

5. (c) **Senior Management**

   Successful firms have shown that analytics usage and deployment is due to the commitment by senior management.

6. (a) **Cleans the data**

   There may be missing values, or values which are not correctly entered at the transaction time.

7. (b) **Build, Test**

   Usually the model is not developed on the entire data. A portion of the data (random) is kept aside for testing. The analyst develops the model (build set) on a portion of the data and tests on another portion (test set).

8. (c) **Machine Learning**

   In this category the analyst does not assume a model before an analysis. The techniques used (such as Machine Learning) will extract a model from the data.

9. (d) **Enormous and Unstructured**

   Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data therefore is data that is enormous and unstructured.

10. (d) Market Basket Analysis

Looks at the shopping basket of its customers to detect patterns. Are there products which appear to be purchased together? Are there products which are never in the same basket? Which sections of the retail store is frequented? Based on the analysis the firm can design marketing actions. These could be adjacent placement of products, product bundling, coupon plans etc.

Unit 11

Business Analytics Techniques

Structure

"It's easy to lie with statistics. It's hard to *tell the truth without statistics."*

- Andrejs Dunkels; Swedish mathematics teacher

11.1 Introduction

Volumes of data is generated globally from various sources like social media, satellite images, GPS, retail, email, videos, and podcasts. It gives opportunity for business teams to leverage the available data using deep analytical tools like data mining, machine learning, data science, and various other statistical analysis.

In the previous unit, we dealt with what is analytics is, its connection to database marketing and usage of database marketing. We have also learnt about big data, data mining, issues in analytics, CRISP-DM Model and four pillars of analytic competition. Other topics dealt in the previous units are, analytics as a competition edge and some typical tasks performed by analysts. Big data computing is an emerging technology used for multidimensional information mining over large-scale infrastructure. The data collected from several scientific explorations and business transactions need different tools for efficient data management, analysis, validation, and visualization. Big data technologies are leading to faster computing and large-scale storage at lower prices. It is enabling large volumes of data preservation and utilization at faster rate. It is predicted that there could be an increase in the digital data by 40 times from 2012 to 2020. The cloud computing technologies are enabling to preserve every bit of the gathered and processed data and providing high availability of storage and computation at affordable price.

In this unit we will be covering the range of data analysis, geospatial intelligence, correlation analysis, regression analysis, multiple linear regressions, logistic regression, factor analysis, Exploratory Factor Analysis (EFA), Principle Factor Analysis (PFA) and Confirmatory Factor Analysis (CFA).

11.2 Objectives

At the end of this unit, you will be able to:

- Discuss how geospatial intelligence is making our life better
- Describe how to maintain balance between creation and consumption of analytics
- Define different mathematical techniques used for big data analytics, like correlation and regression
- Discuss how factor analysis is used to measure relationships between large numbers of variables
- Identify different analysis techniques used in classification

11.3 Make the Last Mile in Data Analysis

Let us understand about the last mile concept data analysis.

The Last Mile

The ‘Last Mile’ is the group of people that delivers the results of the data analysis. The result is given to the businesses so that they can easily understand the trend. It is made up of data analytics to know enough about the business. It requires experienced data analytics professionals who are not afraid to present the results to the CEO.

The last mile group can handle big issues and help in developing and guiding business strategies. For example, Yahoo Mail. When people signup for Yahoo Mail account, first they see news preview module. The news preview module has become popular because it helps in retaining active users. Data analytics professionals noticed that new users like to read news when they read mail. So by adding a news preview window, the return rate increased by 40%. In addition to new users, the other users also liked reading news while looking at their mail.

11.3.1 Geospatial Intelligence Will Make Your Life Better

Geospatial Intelligence means using data about space and time to improve the quality of predictive analysis. For example, smart phone. The smart phone is designed to track the traffic. It shows streets in red and yellow color. It observes the average speed of travel and calculates the aggregate speed to travel, to help us avoid traffic. Therefore, geospatial analytics has become a standard part of life today.

For advertisers also, geospatial intelligence helps in different ways. It helps in enhancing the authenticity of advertisements.

The following are examples of geospatial intelligence.

- Healthcare organizations will be able to predict movements of disease outbreaks over time and adequately prepare for potential epidemics before they occur.
- Police departments can study past geospatial data to see where crimes occurred frequently and understand where and when future crimes are most likely to happen.
- Insurers can incorporate geospatial information into their risk calculations to optimize pricing for known risk factors.

Exhibit 11.1 Explain about IBM geospatial analytics.

Exhibit 11.1: IBM Geospatial Analytics

What can geospatial analytics do for your business?

Formerly IBM PAIRS Geoscope, the geospatial analytics component is a platform specifically designed for massive geospatial-temporal (maps, satellite, weather, drone, IoT) query and analytics services. It frees up data scientists and developers from cumbersome processes that dominate conventional data preparation, providing search-friendly access to a rich, diverse and growing catalog of continually updated geospatial-temporal information.

Gain valuable insights with hundreds of data layers:

The geospatial analytics component offers hundreds of geospatial-temporal data layers, providing unique knowledge that is crucial to maintaining leadership in the era of machine learning and AI. You can expand this knowledge by aligning your organization’s proprietary data for use in combination with the provided geospatial-temporal data, gaining even deeper insights around decisions that matter in a scalable, cost-efficient manner.

Source: IBM Environmental Intelligence Suite - Geospatial Analytics. (n.d.). IBM. https://www.ibm.com/products/environmental-intelligence-suite/geospatial-analytics

Example: Drones for Last-Mile Delivery

Italian city of Milan was focusing on last-mile delivery in supply chain with known value to business and society. Last-mile delivery was a major challenge in supply chain with increased ecommerce operations and supply chain dependency in Urban areas.

Contd….

There were main concerns and factors in last mile logistics like: speed, cost of delivery, and in recent times the environmental and social sustainability factors. Drones were the best choice while considering transport speed and sustainability in the last-mile delivery. In addition, they were electrically powered, reduced noise pollution, which led to reduced environmental impact, and reduced road congestion. Further, during the Covid times, in some deliveries, drones usage lessened direct interaction with end customer, avoiding close contacts.

Source: https://www.researchgate.net/publication/358349469_The_Use_of_Drones_for_Last-Mile_Delivery_A_Numerical_Case_Study_in_Milan_Italy February 2022, Accessed on 12/09/22

11.4 Consumption of Analytics

Consumption of analytics means making analytics consumable in an organization. There are different stages in consumption of analytics.

- Communicate
- Implement
- Measure
- Align incentives
- Develop cognitive repairs

Communication

In this first stage, the business analytics professionals from the core team will meet a wider group of decision makers and the daily consumers of analytics in an organization. It helps create a platform for analytics in the organization.

Implement

Implementation is to get all the ingredients in place to consume analytics successfully. Strong leadership can be the most important trigger in adapting analytics in the organization.

Measure

Measure means testing of consumption. It uses analytics to test itself. A successful business decision can be taken only with a combination of business experience and analytics.

Align Incentives

Successful consumption of analytics leads to creation of structured decision-making processes which is produced by data analysis.

Develop Cognitive Repairs

Creation of business insights based on data and then going and proving it right for all to see, is by far the most effective way to both expose biases and create repairs.

11.4.1 From Creation to Consumption

Various organizations have created analytics but have failed in consumption. Creating analytics does not automatically result in consumption.

The following are some key questions:

- Do you have experience in creating analytics but failed in consumption?
- Do you have problem in maintaining a balance between analytics creation and consumption?

If the answer to any of these questions is 'yes', your organization suffers from creation-consumption gap. Organizations should be able to manage this creation-consumption gap and capitalize analytics as a source of consumptive advantage.

11.4.2 Types of Business Analytics

Big data analytics uses three types of business analytics. They are:

- Descriptive analytics.
- Predictive analytics.
- Prescriptive analytics.

Descriptive analytics describes the past business analytics. It uses SAS and SPSS.

Predictive analytics uses past business analytical information and predicts future outcomes with some degree of likelihood.

Prescriptive analytics uses past business information to direct future activities and to achieve optimal results.

These three techniques are being used for decades, combined with big data, in shifts as follows:

- Using more or all of the data for predictive model.
- Combining analytical models to improve the results
- Using new learnings in predictive models
- Making predictive model close to real time analytics
- Applying predictive models rather than new techniques

Table 11.1 details what each analytic analyses from the data.

Table 11.1: Different Analytics Approaches

Descriptive Analytics	Predictive Analytics	Prescriptive Analytics
➢ What happened? ➢ When did it happen? ➢ What is the problem?	➢ What will happen next? ➢ What if these trends continue? ➢ What if?	➢ What is the best answer? ➢ What is the best outcome?

Contd…..

➢ How often does it happen?		➢ What are the better choices?
It uses ➢ Statistics	It uses ➢ Data mining ➢ Predictive modeling ➢ Forecasting ➢ Simulation	It uses ➢ Constraint-based optimization ➢ Multi-objective optimization ➢ Global optimization

Source: ICFAI Research Center

11.4.3 Correlation Analysis

Correlation is the most useful statistical tool which describes the degree of relationship between two variables. After calculating correlation, we determine the probability of observed correlation by conducting test of significance.

What is Correlation?

If an activity is responsible for another happening, then we can assume that the two will be correlated. Two things that are correlated may not necessarily be related by any definite cause; the reason is a subset of another.

Big data brought with it a great deal of responsibility. Advanced algorithms need to be developed to address questions to analyze vast amounts of data. We can continue to rely on the expertise of data scientists, who have the expertise to ask the relevant and right questions and draw appropriate conclusions.

11.4.4 Regression Analysis

Regression analysis describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age. It involves manipulating some independent variable (example, background music) to see how it influences a dependent variable (example, time spent in a store).

Regression analysis is being used to determine:

- The levels of customer satisfaction.
- The number of support calls received, may be, influenced by the weather forecast given the previous day.
- The neighborhood and size affecting the listing price of houses.

11.4.5 Multiple Linear Regressions

Multiple regression means estimating a single regression with more than one outcome or variable. There are two kinds of variables, independent variable and dependent variable. Once the level of significance between an independent variable and a dependent variable is identified, accurate predictions can be made.

There is a closely resembling term Multivariate Regression. But there is a difference between multiple regression and multivariate regression. In multivariate regression there are more than one dependent variables with different variances (or distributions). The predictor variables may be one or multiple. In multiple regression, there is just one dependent variable, i.e. y. But, the predictor variables or parameters are multiple.

Multiple linear regression explains the relationship between one continuous dependent variable and independent variable. There can be two or more independent variables. Values of independent variables(x) are associated with dependent variables(y).

11.4.6 Logistic Regression

Logistic regression is a statistical method for analyzing a data in which there are one or more independent variables that determine an outcome. Following are the characteristics of logistic regression:

- Logistic regression is a predictive model.
- Logistic regression model does not involve decision trees.
- Logistic regression is used only with two types of target variables:
  - A categorical target variable
  - A continuous target variable

Activity 11.1

You are the Vice President, Marketing of a big retail store chain which has 1200 outlets across India. Your store chain generates very big data every day. Describe how you can use this available big data for descriptive, predictive and prescriptive analysis?

Answer:

Check Your Progress - 1

1. What do you call the group of people who deliver the results of the data analysis?
   a. First mile
   b. Last mile
   c. Data mile
   d. Group mile
   e. Data group mile

2. Using data about which of the following refers to Geospatial Intelligence?
   a. Space
   b. Time
   c. Space and time
   d. Business
   e. People
3. What is the third stage in consumption of analytics?
   a. Communicate
   b. Measure
   c. Develop cognitive repairs
   d. Implement
   e. Align incentives
4. Which of the following is used by descriptive analytics for data analysis?
   a. Data mining
   b. Forecasting
   c. Stimulation
   d. Multi-objective optimization
   e. Statistics
5. Which of the following explains the relationship between one continuous dependent variable and two or more independent variables?
   a. Correlation
   b. Regression analysis
   c. Linear regression
   d. Multi-linear regression
   e. Logistic regression

11.4.7 Factor Analysis

Factor analysis is a tool used to measure the relationship between large numbers of variables. It allows researchers to use psychological scales to measure directly, by collapsing large number of variables.

The main concept of factor analysis is to measure the variables which are associated with a latent (which is not measured directly). For example, people may respond similarly about education, occupation and income which are all associated with socio-economic status which is a latent variable.

In every factor analysis, the number of factors and variables are same, and factors are always listed in the order of variation. Therefore, each factor captures overall variance in the observed variables.

Any factor with an eigenvalue ≥1 can be inferred as having more variance than a single observed variable, because eigenvalue gives the measure of the variance of the observed variables.

Factor Loading

Factor loading means the relationship of each variable under each factor. Here is an example of the output of a simple factor analysis with just six variables and two resulting factors. Table 11.2 presents an example factor analysis.

Table 11.2: Factors Analysis Representation

Variables	Factor 1	Factor 2
Income	0.68	0.13
Education	0.57	0.28
Occupation	0.49	0.21
House value	0.39	0.61
Number of public parks in neighbourhood	0.14	0.55
Number of violent crimes per year in neighbourhood	0.25	0.54

Source: ICFAI Research Center

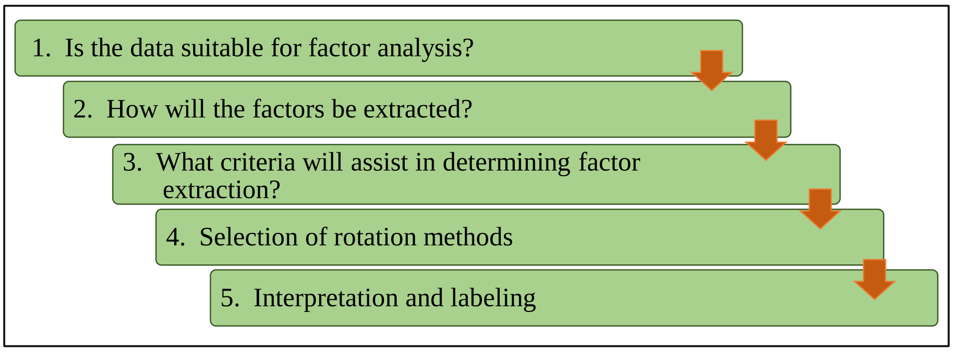
We observe that, the variable with the strongest association (0.68) to the underlying latent variable Factor 1, is income. We can conclude that the variable income has a correlation of 0.68 with Factor 1. For research fields, this would be considered a strong association for a factor analysis.

11.4.8 Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) is a statistical method used to uncover the underlying structure of a relatively large set of variables. Factor analysis uses EFA technique to identify the underlying relationships between measured variables. EFA is a complex statistical approach involving many options. The following are the five steps involved in EFA.

Figure 11.1 presents five step exploratory factor analysis

Figure 11.1: Five Step Exploratory Factor Analysis Protocol


1. Is the data suitable for factor analysis?
2. How will the factors be extracted?
3. What criteria will assist in determining factor extraction?
4. Selection of rotation methods
5. Interpretation and labeling

Source: ICFAI Research Center

11.4.9 Principal Factor Analysis (PFA)

Principal factor analysis is used for factor extraction. This is the first phase of EFA (Exploratory Factor Analysis). In this, factor weights are calculated to produce the maximum possible variance until further meaningful variance is left are as follows;

- PFA is computationally quicker and requires a few more resources than factor analysis.
- It can produce smaller results.
- The results of PFA are inaccurate results.
- It computes factor source from factor analysis.

11.4.10 Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis (CFA) and Exploratory Factor Analysis (EFA) are similar techniques. Confirmatory factor analysis is a tool that is used to confirm or reject the measurement theory. In this, the researchers can specify the number of factors required in the data, and the relationship between a measured variable and a latent variable.

It is a multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs.

The following steps are involved in CFA.

1. **Defining individual construct:**

   First, we have to define the individual construct. This involves a pretest to evaluate the construct items, and a confirmatory test of the measurement.

2. **Developing the overall measurement model theory:**

   We should consider the concept of uni-dimensionality between construct error variance and within-construct error variance.

3. **Designing a study to produce empirical results:**

   The measurement model must be specified. There are two methods, the first is rank condition and the second is order condition.

4. **Assessing the measurement model validity:**

   In this, the theoretical measurement model is compared with the reality model to see how well the data fits. Chi-square test and other goodness of fit statistics like RMR, GFI, NFI, RMSEA, SIC, BIC, etc., are some key indicators that help in measuring the model validity.

Example: Data-driven building maintenance at Finland

Ville Kautto was the Maintenance Manager of University Properties of Finland (SYK), as well as to many other universities in Finland. SYK strived to be the best service provider for campuses in Europe.

Contd….

To understand detailed status of their equipment, usage of the premises, they moved to digitalization, which helped constant monitoring of equipment, evaluation of indoor climate centrally, and real time addressing of the needs. SYK joined with Siemens and revamped the maintenance approach at Tampere University. They realised that, visual inspections had reduced by 70% by use of data analytics.

Digital brought a change in the campus with sensors for measurement, collection and formatting, smart algorithms, infinite cloud computing power, accurate monitoring and demand-driven work. Data analytics experts searched and analyzed for optimization. They used "value-hacking", "design thinking" to integrate forecasted weather data for the BMS (Building Management System). The data from recorded ground temperature, weather forecasts were used to attend to melt snow from campus walkways and avoid heating them. This data-driven approach helped to focus on roofs also when upcoming snowstorm was detected. This led to 50% reduced user complaints, 57% energy efficiency, 70% of data analytics usage instead of visual inspections.

Source: https://new.siemens.com/global/en/company/stories/infrastructure/2020/data-driven-maintenance-project-finland.html November 2020, Accessed 29/08/22

11.5 Classification

Find below various analysis approaches-

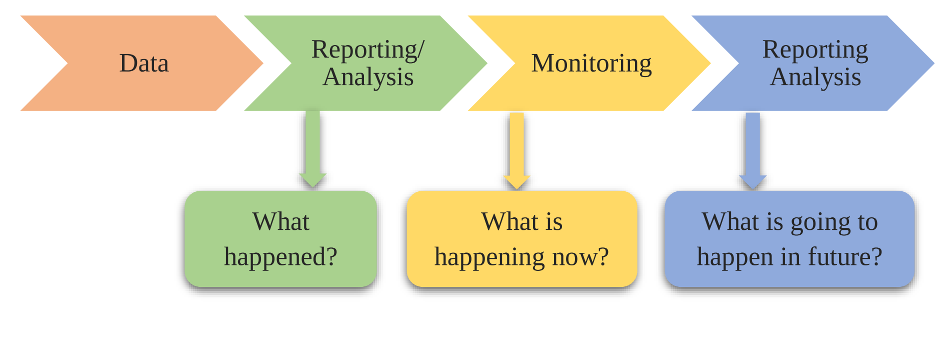
11.5.1 Predictive Analysis

Predictive analysis means making predictions about future events. It uses many techniques like data mining, statistics and artificial intelligence to analyze the present data to generate predictions about future and identifies risks and opportunities for future.

Predictive analysis makes organizations more proactive and forward looking. It provides decision options to get benefits from the predictions.

Figure 11.2 shows predictive analysis.

Figure 11.2: Predictive Analysis Value-Chain



Source: ICFAI Research Center

Predictive Analysis Process

- **Define Project**

  Predictive analysis defines the business objectives, project outcomes and identifies data sets for that project.

- **Data Collection**

  Predictive analysis collects data from multiple sources for analysis and also for customer interaction.

- **Data Analysis**

  Predictive analysis analyzes the data using the process of inspecting, transforming and modeling to discover useful information.

- **Statistics**

  Predictive analysis uses statistical methods to validate the data assumptions and hypothesis.

- **Modeling**

  Predictive analysis creates models for future so that the management can choose the best model for optimizing the assumptions.

- **Deployment**

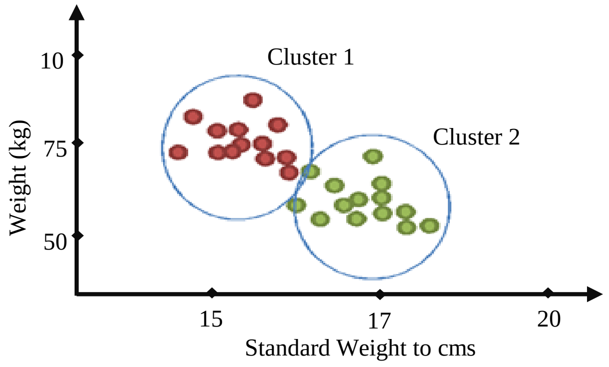
  Predictive analysis deploys the analytical results into decision-making process to get better results.

11.5.2 Cluster Analysis

Cluster means collections of similar objects into a group. The objects in a cluster have similar properties. Cluster analysis means organizing the group of similar objects. Cluster analysis is used for descriptive statistics. The main aim of clustering is to make best criterion which is independent.

Figure 11.3 presents cluster approach.

Figure 11.3: Collection of Similar Objects in a Cluster



Source: ICFAI Research Center

The following are the different types of clusters:

- **Exclusive Clustering**

  Exclusive clustering groups objects in an exclusive way so that the objects belong to one definite cluster.

- **Overlapping Clustering**

  Overlapping clustering uses fuzzy logic so that one object may belong to one or more clusters with different degree of membership.

- **Hierarchical Clustering**

  There are two types of hierarchical clustering:

  ➢ **Agglomerative Clustering**

  It is a bottom-top version. Initially, each object is treated as a cluster and after few iterations like union between the nearest clusters, it reaches the final cluster.

  ➢ **Divisive Clustering**

  It is a top-down version. It starts with a cluster, starts with all objects, and later clusters are split into smaller clusters.

- **Probabilistic Clustering**

  Probabilistic clustering uses completely probabilistic approach to group the objects into clusters. For example, mixture of Gaussian.

11.5.3 Association Analysis

Day-to-day operations in business accumulates large quantities of data. For example, at a grocery store at each counter, huge data of customer purchase is collected every day. The following Table 11.4 illustrates such data.

Table 11.4: Market Basket Items

TID	Items
1.	Bread, milk
2.	Bread, diapers, beer, eggs
3.	Milk, diapers, beer, cola
4.	Bread, milk, diapers, beer
5.	Bread, milk, diapers, cola

Source: ICFAI Research Center

TID is the unique transaction id of each customer. The purchase behavior of the customer can be analyzed by the retailors, and association relationship will be given to the frequent items.

For Example:

There is a strong relationship between the sales of diapers and beer, because the customers who purchase diapers also purchase beer.

Diapers --------Association------→ Beer

The retailers can use this association to get opportunities for cross-selling their products to the customer.

Example: Exploratory Data Analysis usage in risk minimization

One of the loans providing companies had inadequate or missing credit history and decided to move on to analytics for disbursement of loans. The suitable analytics approach was Exploratory Data Analysis (EDA), which mined the data to assess applicants with proper repaying capability, and minimize the risk. When the company received a loan application, the four decisions could be: Approved, Cancelled, Refused, and Unused offer. The major goal was: do not reject loan to applicants with capability to repay. Exploratory Data Analysis and use of Machine Learning algorithms could successfully classify fraudulent and positive cases.

Source: https://www.analyticsvidhya.com/blog/2022/03/exploratory-data-analysis-eda-credit-card-fraud-detection-case-study/ April 2022, Accessed on 29/08/22

Activity 11.2

You are the Area Head for a business marketing research group. You are doing study on behavior of retail buyers across five major retail outlet brands like Heritage, Ratnadeep, Reliance, Ghansyam and DMart at Hyderabad. How best you can take benefit of Cluster analysis in this activity?

Answer:

<u>Check Your Progress - 2</u>

6. Factor analysis measures the variables which are associated with which of the following?

   a. Cluster

   b. Object

   c. Products

   d. Latent

   e. People

7. Which of the following techniques is used to identify the underlying relationships between measured variables?
   a. Exploratory factor analysis (EFA)
   b. Principal factor analysis (PFA)
   c. Confirmatory factor analysis (CFA)
   d. Predictive analysis
   e. Cluster analysis
8. Collections of which of the following objects into groups is called cluster?
   a. Different
   b. Similar
   c. Principal
   d. Exclusive
   e. Probabilistic
9. Which type of clustering uses fuzzy logic so that one object may belong to one or more clusters with different degree of membership?
   a. Exclusive clustering
   b. Overlapping clustering
   c. Hierarchical clustering
   d. Probabilistic clustering
   e. Probabilistic clustering
10. What do you call the analysis where customers who purchase diapers also purchase beer (Diapers -----------→ beer)?
    a. Predictive analysis
    b. Cluster analysis
    c. Association analysis
    d. Descriptive analysis
    e. Perspective analysis

11.6 Summary

- Big data organizes and extracts the information from the rapidly growing, large volume of variety of forms, and frequently changing data sets collected from multiple and autonomous sources in the minimal possible time, using several statistical techniques.
- Big data technology is changing the present traditional data bases with effective data organization.
- Big data technologies are accelerating in several areas of business, science and engineering to solve problems.

- In this unit, we discussed different technologies used for data analytics. We also discussed cluster analysis and different types of clustering technologies that support big data technology

11.7 Glossary

Confirmatory Factor Analysis (CFA): Confirmatory factor analysis (CFA) is a tool that is used to confirm or reject the measurement theory.

Eigenvalue: Eigenvalues are a special set of scalars associated with a linear system of equation (i.e., a matrix equation) that are sometimes also known as characteristic roots or characteristic values.

Exploratory Factor Analysis (EFA): Exploratory factor analysis (EFA) is a statistical method used to disclose the underlying relational structure of relatively large set of variables.

Latent: Latent means something that is capable of becoming active or at hand but has not yet achieved that state.

Principle Factor Analysis (PFA): Principal factor analysis is used for factor extraction. In this factor, weights are calculated to produce the maximum possible variance until further meaningful variance is left.

Socio-Economic Status (SES): SES is an economic and sociologically combined total measure of a person's work experience, and of an individual's or family's economic and social position, in relation to others, based on income, education and occupation.

11.8 Self-Assessment Test

1. Describe how Geospatial Intelligence will make your life better.
2. What is Correlation? How it is useful in big data analytics?
3. Analyze the consumption of analytics from creation to consumption.
4. What is regression? What are the different regression analyses available?
5. What is factor analysis? What are the different factor analyses that support big data analysis?

11.9 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.

5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition

6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

11.10 Answers to Check Your Progress Questions

1. (b) Last mile

‘Last Mile’ is the group of people who deliver the results of the data analysis. This result is given to the businesses so that they can easily understand the trend.

2. (c) Space and time

Geospatial Intelligence means using data about space and time to improve the quality of predictive analysis.

3. (b) Measure

The third stage in consumption of analytics is ‘Measure’. Measure means testing the consumption. It uses analytics to test itself. A successful business decision can be taken only with the combination of business experience and analytics.

4. (e) Statistics

Descriptive analytics describes the past business analytics. It uses SAS and SPSS for descriptive statistics.

5. (d) Multi-linear regression

Multivariate regression means estimating a single regression with more than one outcome variable. It explains the relationship between one continuous dependent variable and two or more independent variables.

6. (d) Latent

The main concept of factor analysis is to measure the variables which are associated with a latent (which is not measured directly).

7. (a) Exploratory factor analysis (EFA)

Exploratory factor analysis (EFA) is a statistical method used to uncover the underlying structure of a relatively large set of variables. Factor analysis uses EFA technique to identify the underlying relationships between measured variables.

8. (b) **Similar**

Cluster means collection of similar objects into a group. The objects in a cluster have similar properties. Cluster analysis means organizing the group of similar objects.

9. (b) **Overlapping clustering**

Overlapping clustering uses fuzzy logic. Here, one object may belong to one or more clusters, and may have different degree of membership.

10. (c) **Association analysis**

The purchase behavior of the customer can be analyzed by the retailors and the association relationship will be given to the frequent items. The customers who purchase diapers also purchase beer.

Diapers --------Association------→ beer

Unit 12

Data Visualization and Modelling

Structure

"Measure the most you can and show the least you can."

- Danique Roefs

12.1 Introduction

While the data collected is of huge volumes, the presentation of data visually by making it simple and effective will lead to better business situations and decision making by management.

In the previous unit, we dealt with certain aspects go business analytics covering: correlation analysis, regression analysis, multiple linear regressions, logistic regression, factor analysis, exploratory factor analysis (EFA), principal factor analysis (PFA), and confirmatory factor analysis (CFA). Today we are at an inflection point at which we have computational and intellectual power by big data analytics. It creates a friction-free environment for your business and gives good financial results. Predictive analysis is the well-entrenched assumptions that have realized significant and quantifiable business value.

In this unit we will be to discussing different business analytic techniques like visualizing data, 360 degrees modeling, get scrappy etc. Visualizing data is a technique to facilitate patterns in the data and make it more consumable. The big data scientists are using whole data to create a model in order to increase the accuracy of the model.

12.2 Objectives

At the end of this unit, you will be able to:

- Explain data visualization and how it makes data more consumable
- Differentiate usage of sample and population in data analysis
- Discuss the concept of 360 degrees modeling and its advantages
- Discuss how big data technologies enable faster data processing
- Discuss the new technologies in big data analysis

12.3 Visualizing: How to make it Consumable?

Visualizing data is a technique used to present the data in patterns to make it more consumable. Generally, business analytics use graphs, charts and dashboards to represent the data in comprehensive format.

It uses the intents such as:

- ***Describing:*** It explains the basic meaning of the data.
- ***Reporting:*** It generates summary of the data.
- ***Observing:*** It views the data to observe patterns over a period of time.
- ***Discovering:*** It interacts with the data to understand the relationships in data.

These intents are useful for current volume of information. However, these intents fail, if the data expands substantially. Showing billion points of data on a chart, graph, or dashboard is impractical. To create new and dynamic visualizations new technologies and tools emerged. These new technologies visualize the data that exceed the standard charts, graphs, and dashboards.

The people who create these new visualizations are called ‘data artisans’. These data artisans are skilled in science, design and art. The Exhibit 12.1 discusses about data artisans.

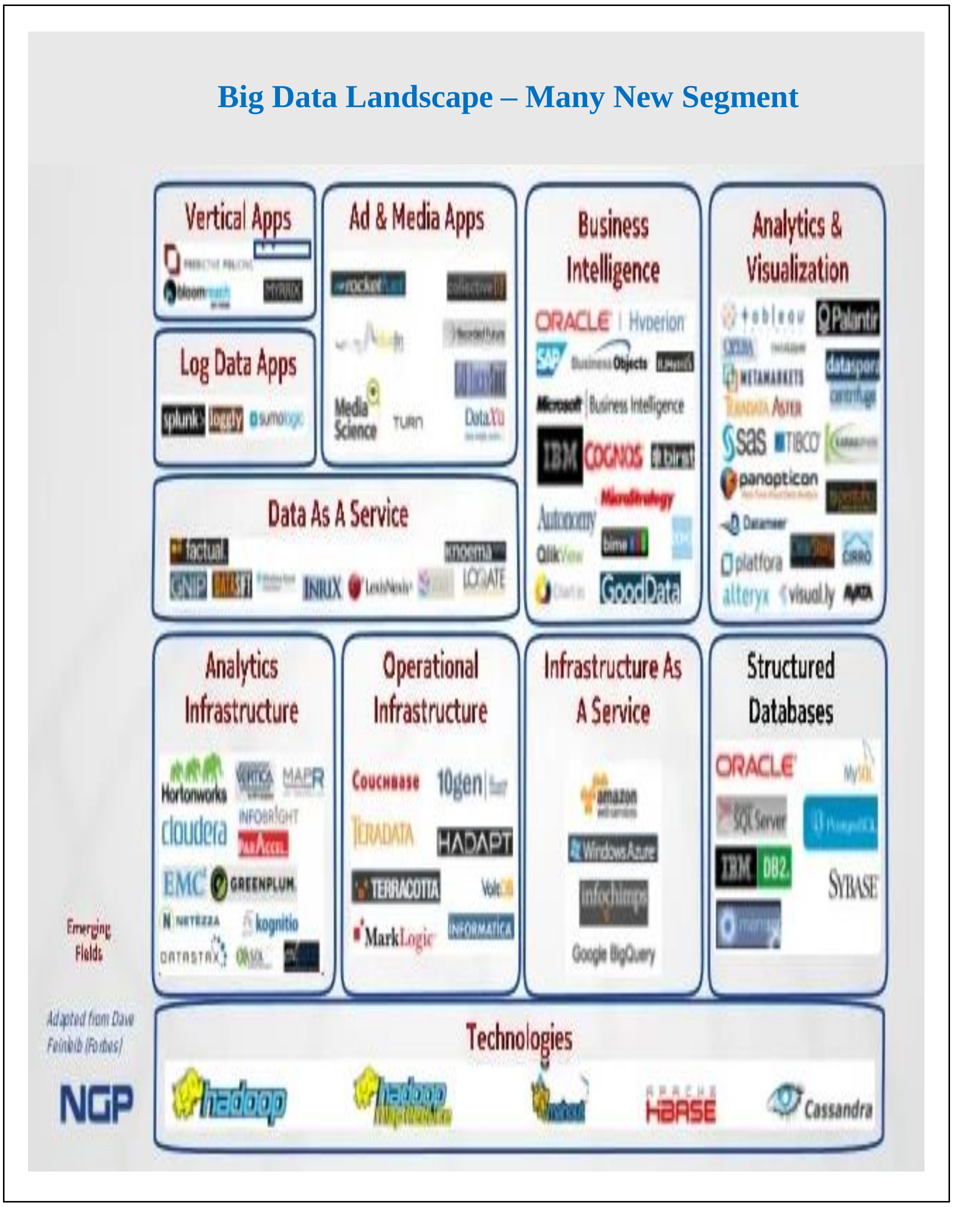
The data artisans are using different dimensions and attributes to represent the data. They are like

- Spatial, Geospatial (positions, directions etc.)
- Temporal (state, phase etc.)

- Scale, Granularity (count, weight, size etc.)
- Resources (temperature, energy etc.)
- Relativity, Proximity
- Value, Priority
- Constraints

Figure 12.1 displays big data landscape of.

Figure 12.1: Big Data Landscape Visualization



Source: Chakraborty, M. (2021, March 5). 6 Factors which will change the Big data landscape 2020. Tech Blogger. https://contenteratechspace.com/blogs/big-data-landscape-2020/

Example: Visual Data Insights at a Healthcare Provider

The CEO of a major health care provider was keen on data analytics of the data generated from patients, service providers, and all the health care systems in place at a given time. However, presentation of the collected data, for the digestion of the viewers was a big challenge. He was looking at annotating of insights.

He looked at adding visual elements as a layer, to bring clarity and detail the most crucial details in an elegant way. Recommendations were given to pick up right classic charts: bar graph, line graph, and pie chart etc., for bringing visualization to the data. Subsequently the five annotation strategies recommended included: Highlight data, Label data, Bracket data, Delineate data and Explode data. Some other tips included: color, a math, or bubbles.

While sharing critical data insights, he was advised to ensure the approach of 'what you see is what you get.'

Source: https://sloanreview.mit.edu/article/make-your-data-insights-visually-consumable/ October 2020, Accessed on 17/09/22

12.4 Organizations are Using Data Visualization as a Way to Take Immediate Action

The companies are collecting billion bytes of data a day and generating processes for analytics and reporting. Product managers, analyst and project heads analyze hundreds of millions of data to understand the user dynamics and problems.

The data visualization is making big data analytics iterative. It is also reducing cycle time of big data analytics so that immediate action can be taken.

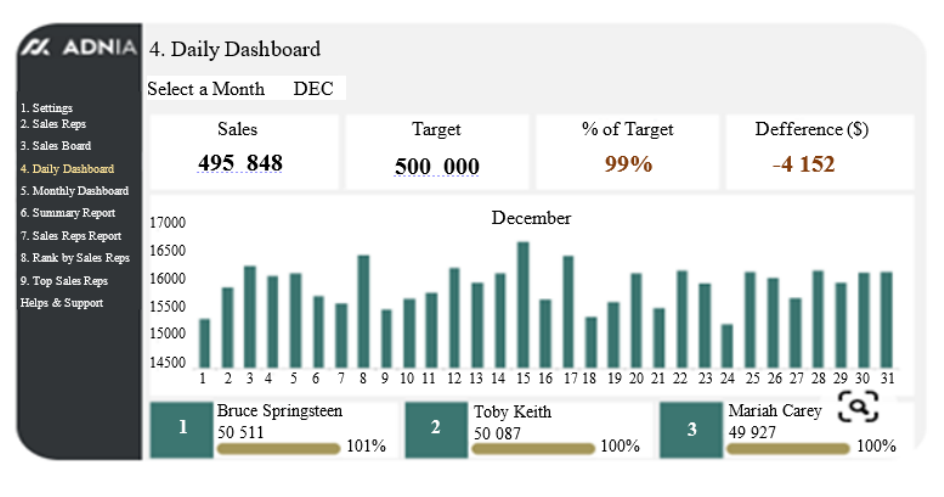
Big data visualization is evolving and commercial vendors are using popular open-source project application software.

The following are some open-source projects:

- Qlikview (www.qlikview.com)
- Tableau (www.tableausoftware.com)
- Microstrategy (www.microstrategy.com)
- SAS – www.sas.com
- Cubism (a plug-in for D3 for visualizing time series) – http://cubsim.com
- Arbor JS, a java-based graph library (http://arborjs.org)
- Many Eyes - IBM Research: Data Visualization Tools

Figure 12.2 presents sales dashboard.

Figure 12.2: Presents Sales Dashboard



ADNIA

4. Daily Dashboard

Select a Month DEC

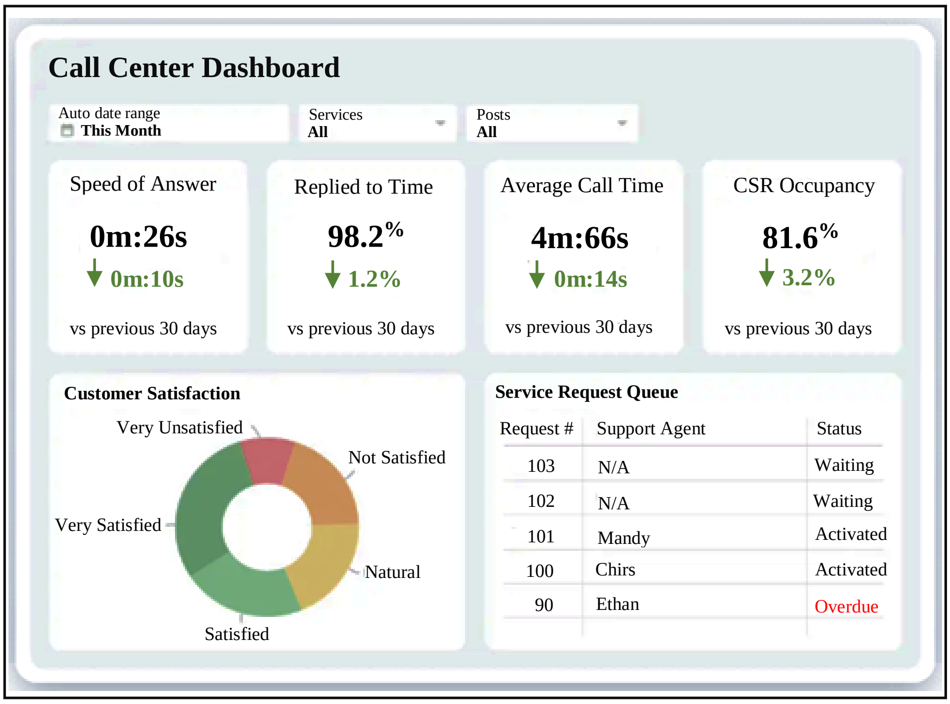
Sales	Target	% of Target	Defference ($)
495 848	500 000	99%	-4 152

1	Bruce Springsteen 50 511	101%
2	Toby Keith 50 087	100%
3	Mariah Carey 49 927	100%

Source: https://www.pinterest.ca/pin/570127634069335621/?autologin=true

Figure 12.3 represents a call center dashboard.

Figure 12.3: Call Center Dashboard



Call Center Dashboard

Auto date range: This Month
Services: All
Posts: All

Speed of Answer	Replied to Time	Average Call Time	CSR Occupancy
0m:26s	98.2%	4m:66s	81.6%
↓ 0m:10s	↓ 1.2%	↓ 0m:14s	↓ 3.2%
vs previous 30 days	vs previous 30 days	vs previous 30 days	vs previous 30 days

Service Request Queue

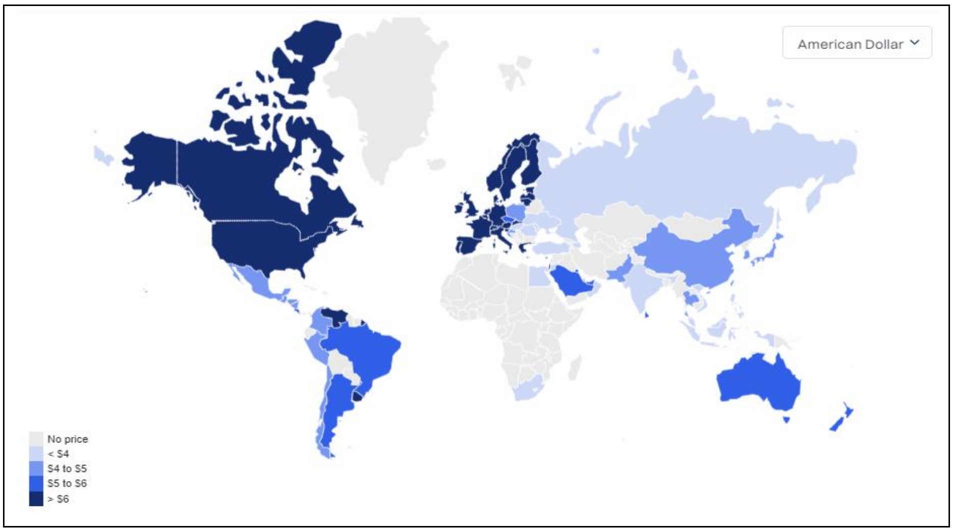
Request #	Support Agent	Status
103	N/A	Waiting
102	N/A	Waiting
101	Mandy	Activated
100	Chirs	Activated
90	Ethan	Overdue

Source: https://www.klipfolio.com/resources/dashboard-examples/call-center

The data visualizations in above figures show dashboards that track sales metrics and call center metrics to allow managers to monitor performance of their teams. These visual displays of data simplify data analysis and present information that helps understand team performance at a glance. The filters on the dashboards can be used to select time range, specific services etc.

Figure 12.4 presents the Big Mac index.

Figure 12.4: Presents the Big Mac Index as Visual

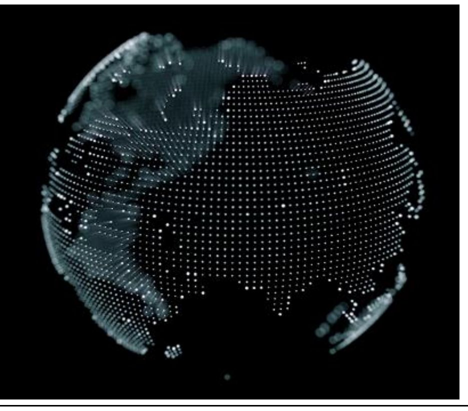


Source: https://hellosafe.ca/en/blog/big-mac-index

Big Mac Index visualization compares the price of a Big Mac burger at McDonald's in different countries. Big Mac burger represents a standardized product that includes input costs from various areas, such as agricultural commodities (beef, bread, lettuce, cheese), labor (blue and white collar workers), advertising, rent and real estate costs, transportation etc., and therefore, is representative of the general state of the local economy.

Figure 12.5 presents a Google data art.

Figure 12.5: Google Data Art



Source: https://in.pinterest.com/pin/374432156530624883/

This Globe is created by the Google Data Arts Team. It is an open platform for geographic data visualization. Users can copy the code, create their own globes, and add their own data. It requires latitude, longitude, and the magnitude related to the location. The user generated visualizations can then be featured on the project's website.

Example: Data Exploration and Visualization at Netflix

Netflix was a famous media and video streaming platform, with 10000 movies/ TV shows, 222 million subscribers globally. The challenge was, ensuring the presentation of data like cast, directors, ratings, release year, duration, etc. of all the movies and TV shows details. They wanted these to be presented in visual way for easy selection by the customer. The data structures consisted of: Show_id, Type, Title, Director, Cast, Country, Date added, Release year, Rating, Duration, and Description. Netflix picked up top 10 genres and drilled deeper for insights on user's interests. They plotted a graph to locate the top genres. Netflix wanted to generate insights on type of shows, and growing business in multiple countries. For importing relevant libraries to mine the dataset, they used numpy, pandas, for analytics and matplotlib, seaborn for visualization. The distribution was skewed relative to the type of content present.

Source: https://medium.datadriveninvestor.com/netflix-data-exploration-and-visualization-1d270234c2d4 case study July 2022, Accessed on 17/09/22

Activity 12.1

From the various Data visualization projects listed herein, pickup one project and give complete details about objective, processes and benefits.

12.5 Moving from Sampling to Using All the Data

For decades, we have been using the statistical approach called 'sample', which is a subset of data. A sample is taken from the whole data to validate by conducting a test. If the sample is validated, then we consider that the whole data is validated to create a model.

Big data scientists are using more or all of the data instead of a sample so that it increases the accuracy of the model. It also allows the scientists to introduce additional predictor variables in the model and identifies the trends in historical data extraction.

Example: Event Detection Ensemble Model

On 4 August, 2020 an enormous non-nuclear blast occurred at the Beirut Port. Identifying what happened 'during and after emergency' events was essential and crucial for protection of humans' life, assess damages to the environment, losses to infrastructures, and all allied financial consequences.

Contd….

The blast killed around 200 people, wounded over 6000, made 300,000 homeless, and $15 billion in destruction of property. Machine learning based models were ensembled for detection of the emergency events.

Instead of using samples, full set of snaps at the location were classified using Convolutional Neural Network(CNN), and transfer learning approach; using the pre-trained model(ResNet50) for feature extraction. Tracing from the Snapchat map on exact location of the emergency event in Beirut, 50,244 tweets were collected.

Apache Spark streaming was used to integrate the trained models. The results proved about the ensemble approach which achieved 99.87%, very high accuracy. Future ensembling models would cover complete data collected from vivid social network platforms, exploring IoT sensors and satellite imagery sources, instead of using any sampling methods.

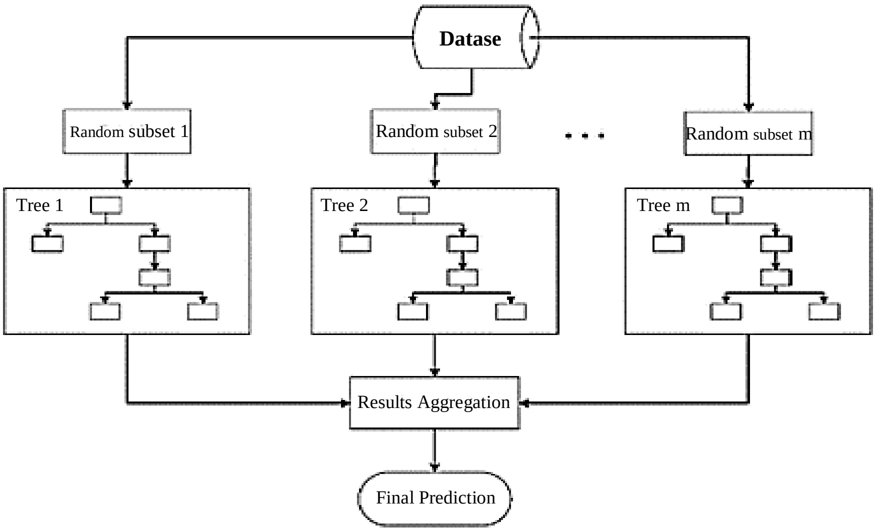
Source: https://www.mdpi.com/2504-2289/6/2/42/pdf-vor April 2022, Accessed on 17/09/22

12.6 Thinking Outside the Box

By thinking outside the box, big data scientists are generating better results for their enterprise. They are using multiple techniques to create models to meet business objectives. By using multiple techniques, they are creating multiple models, which are called ‘ensemble models’.

The scientists chain the techniques together to provide the best result. They are using predictive model algorithms and simulation techniques to evaluate millions of scenarios. Then they are applying optimization technique to maximize the model output. Exhibit 12.3 present ensemble methods.

Exhibit 12.3: Ensemble Methods: Bagging



Contd….

- Ensemble Methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models, M_1, M_2,…, $M_{k,}$ with the aim of creating an improved model M*
- Popular Ensemble Methods
  - Bagging: Averaging the prediction over a collection of classifiers
  - Booting: Weighted vote with a collection of classifiers
  - Ensemble: Combining a set of heterogeneous classifiers

Source: https://hevodata.com/learn/ensemble-data-mining/#:~:text=Ensemble%20Data%20Mining%20is%20the,to%20lessen%20the%20prediction%20error.

<u>Check Your Progress - 1</u>

1. Which of the following intent interacts with the data to understand relationship in data?

   a. Describing

   b. Reporting

   c. Observing

   d. Discovering

   e. Analyzing

2. Which of the following terms refer to the people creating new visualizations?

   a. Designers

   b. Artists

   c. Artisans

   d. Creators

   e. Producers

3. Which of the following is making big data analytics iterative?

   a. Data scale

   b. Data resources

   c. Data priority

   d. Data visualization

   e. Data state

4. Which of the following is used by big data scientists to conduct tests so that it increases the accuracy of the model?

   a. Sample data

   b. More data

   c. All of the data

   d. More or all of the data

   e. None of the above

5. Which of the following term refers to big data scientists generating better results using multiple models?

   a. Sample models

   b. Assemble models

   c. Ensemble models

   d. Artisan model

   e. Complex models

12.7 360^0 Modeling

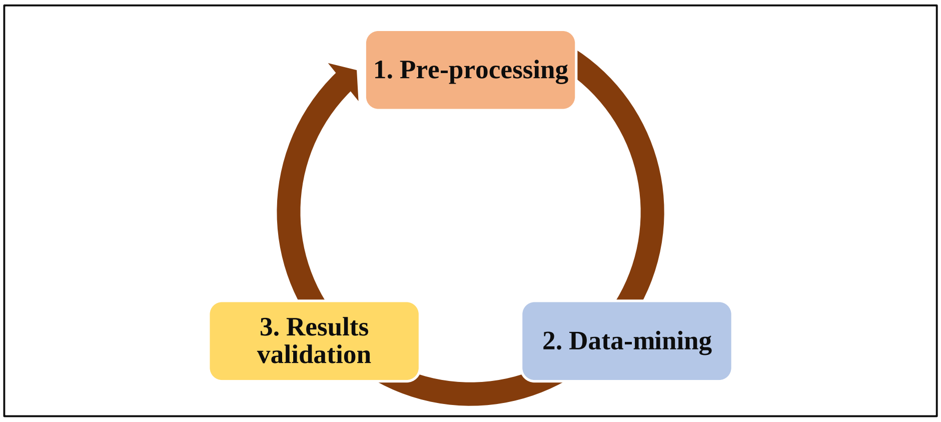
The models deployed periodically and re-evaluated by the creators is called as 360 degrees modeling. The frequency of re-evaluation is very important and varies tremendously.

The big data environment demands that models are continuously re-evaluated to obtain the additional features in the models.

360 degrees modeling creates a closed loop where learnings are incorporated by the data scientists. The learning techniques can be employed and adapted based on the new data.

Figure 12.6 gives 360 degree modeling process.

Figure 12.6: 360^0 Modeling Process



Source: ICFAI Research Center

12.8 Need for Speed

Enterprises are moving from manual process to automated process for efficiency. At the same time the enterprisers are trying for faster data processing in this competitive world, because data is analyzed quickly as it is consumed.

Big data technologies enable faster data processing and turning the data into insights. Big data analytics and its insights will always be a part of forecasting. Fast data will enable businesses to make decisions in real time as it flows into the organization.

For example, the sensors on the car send information to the automotive manufacturer and the manufacturer anticipates issues using this information by sending alert message to the driver.

Example: Amazon experiences with Big Data

Amazon, a well-known e-Commerce platform, had different services like: Amazon Pay, Amazon Pantry, Amazon Web Services (AWS) etc. It compiled huge amount of data from various sources and

leveraged on the collected data using big data related analytics and, recommendation engine. Amazon was a leader in collecting data at faster rates about customers, store and access to analyze on their spending patterns, and targeted marketing, using predictive analytics.

Amazon, with camera, and speakers (Echo and Echo Show) and voice commands helped in knowing weather reports, news, as well as ordering shampoo. These uploaded voice files, dynamically helped enhance the Alexa experience, by getting better at speech recognition, and message processing with higher accuracy.

Amazon used a ‘comprehensive, Collaborative Filtering Engine’ (CFE), and behavioral analytics. The information from earlier purchased items, shopping cart items, wish list was used to prompt the customer for additional products for choice. An estimated 35% of annual sales were attributed to this feature. Amazon came up with One-Click ordering, a patented feature, which was automatically enabled after first order. Amazon's ‘anticipatory delivery model’ predicted items, a customer was likely to order, and shipped them off to a neighborhood stockroom.

Amazon integrated Kindle function with social networking service of around 25 million users, helped share highlighted words among peers, and discussed the book. This also helped recommend ebooks to customers and in parallel enhance reading experience.

Source: https://www.analyticssteps.com/blogs/how-amazon-uses-big-data May 2021, Accessed on 17/09/22

12.9 Let’s Get Scrappy

Get Scrappy supports with the courage and commitment to pursue goals, especially when there is no evidence that one can succeed.

Today there are mountains of enterprise problems to solve. The increase in revenue and cost-savings are significant enough while solving the problems.

The big data analytics thinks that rather than spending time in building new techniques, they need to harness best to get scrappy, to solve real world problems. They should focus on recognition of accomplishing something tangible.

For example, when I need my car cleaned, I can pay someone to clean it but I feel so much better when I clean it myself. The fruits of my labor are visible and tangible. Exhibit 12.5 shares the scrappy project management checklist.

Exhibit 12.5: Scrappy Project Management Checklist

Scrappy project management says:

- Be completely obsessed with the “CUSTOMER”
- Provide shared, measurable, challenging & believable GOALS as clear as sunlight
- Engage in effective, vociferous & unrelenting COMMUNICATION with all stakeholders
- Ensure that ROLES & RESPONSIBILITIES are unmistakably understood and agreed upon by all
- Create viable PLANS & SCHEDULES that enjoy the team’s hearty commitment
- Mitigate big, hairy, abominable RISKS & implement innovative ACCELERATORS
- PRIORITIZE ruthlessly, choosing between heart, lungs & kidney if necessary
- Anticipate and accommodate necessary & inevitable CHANGE
- Challenge ASSUMPTIONS & BELIEFS, especially insidious self-imposed limitations
- Manage the EXPECTATIONS of all stakeholders: under-promise & over-deliver
- LEARN from experience. Make new and more exciting mistakes each time!
- ATTITUDE OF GRATITUDE. CELEBRATE project success…and some failures, too!

Source: https://wiefling.com/wp-content/uploads/2009/08/scrappy_checklist1.jpg

12.10 What Technology is Available?

Towards big data analytics many software players are working in an open-source environment. They are IBM, SAS, SPSS, KXEN, Matlab, Statsoft, Tableau, Pentaho etc.

Most of these tools have high speed connectors to move data between Hadoop and their tools. The main objective of big data is to keep the data in one place and avoid the data movement. Therefore, the vendors are developing a strategy to keep data in place and move their analytics processing to the data.

There are new commercial vendors for big data analytics like:

- *Karmasphere* (https://karmasphere.com)

  It is a tool for data exploration and visualization.
- *Datameer* (http://www.datameer.com)

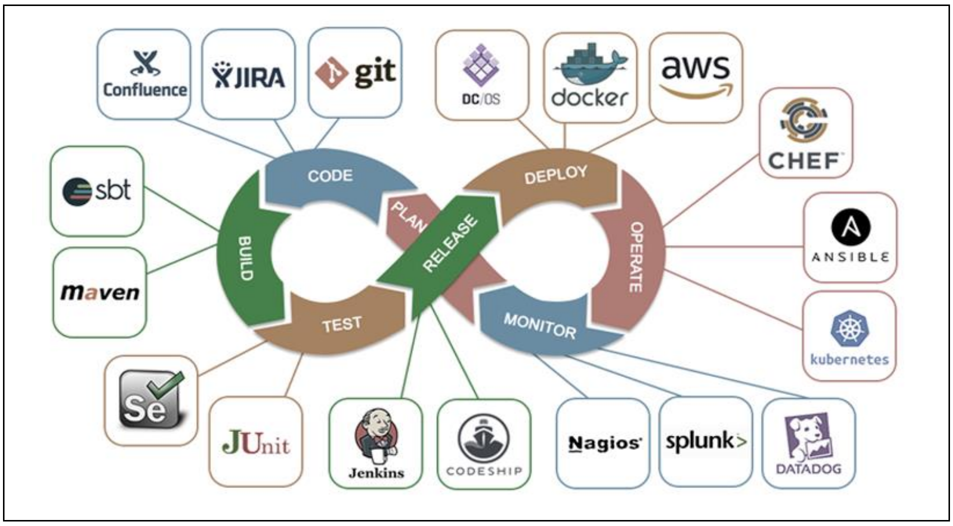
  It is a spreadsheet presentation tool.
- *Alpine Data Miner* (http://www.alpinedatalabs.com)

  It is a cross-platform analytic workbench.
- *R* (http://cran.r-project.org)

  It is a statistical tool which explores and build models.

Figure 12.7 presents The Ultimate List of Open Source DevOps Tools.

Figure 12.7: The Ultimate List of Open Source DevOps Tools



Source: https://www.techtrainees.com/tag/the-ultimate-list-of-open-source-devops-tools/

12.11 Moving from Beyond the Tools to Analytics Applications

In order to capitalize on big data value the big data apps have started to emerge. The horizontal big data apps (machine log analytics) and vertical big data apps (telecommunications analytics) are emerging.

These emerging techniques are designed to solve specific business problems, which incorporate deeper and more complex prescriptive analytics.

The following are the Top Emerging Technologies-

- Fuel-cell vehicles: Cars that run on hydro zed
- Next generation robotics: Rolling away from the production line
- Recyclable thermoset plastics: A new kind of plastic to cut landfill waste
- Precise genetic-engineering techniques: A breakthrough offers better crops with less controversy
- Additive manufacturing: Making things from printable organs to intelligent clothes
- Emergent artificial intelligence: What happens when a computer can learn on the job?
- Distributed manufacturing: The factory of the future is online—and on your doorstep
- Neuromorphic technology: Computer chips that mimic the human brain

Example: Woolworth's analytics for data driven retailing

Woolworths Group, Australia's largest private sector employer, was food and daily needs retailer, and had around 1,400 stores, and 180,000 employees.

Woolworth, wanted to enhance data accuracy, achieve speed, ensure smarter, faster decision-making, mine the big data, to gain competitive edge. They associated with TCS as a strategic partner in this endeavor.

They jointly embarked on cloud transformation with incremental and agile approach. They established a common data platform for all major business functions and supply chain activities on Google Cloud. This helped in reliable, integrated, and AI-driven retail experiences. They built company-wide single source data lake with self-service analytics, for creating new value. Minute insights into challenges and opportunities were made available in quicker and cost-effective way.

90% of the company-wide data was onto Google Cloud, which led to on-demand access to information, detailed and three times faster self-service analytics, descriptive and prescriptive analytics. Disaster recovery was reduced to two hours from five days. .

Woolworths could adjust prices, in each stores dependent on customer groups, could reliably and accurately predict sales patterns for meeting sustainability goals, and improve overall supply chain performance. In turn, it made seamless, enriched experiences for employee as well customer, and gave Woolworths the required competitive edge in retail environment.

Source: https://www.tcs.com/woolworths-data-driven-retail-enhance-customer-experience 2022, Accessed on 17/09/22

Activity 12.2

You are the L and D head (learning and development) of an organization with 1.89 lakh people. Various training programs, nominations to seminars and permitting higher studies are part of the HR process. How best you can use 360 degree modeling for the benefit of assessing effectiveness, knowledge management repository, etc.

Check Your Progress-2

6. Which of the following is done periodically in 360 degrees modelling?
   a. Processed
   b. Refreshed
   c. Re-evaluated
   d. Redesigned
   e. Reconfigured

7. Which of the following will enable businesses to make decisions in real time?
   a. Data base
   b. Big data
   c. Data mart
   d. Fast data
   e. Slow data

8. Which of the following enables you to pursue your goals even though there is no evidence that you can succeed?
   a. Get scrappy
   b. Get model
   c. Get solution
   d. Get success
   e. Get frequency

9. Which of the following is an open source statistical tool that explores and build models?
   a. Karmasphere
   b. Datameer
   c. R
   d. Alpine Data Miner
   e. Apache

10. Which of the following emerging technology that makes computer chips that mimic the human brain?
    a. Emergent artificial intelligence
    b. Precise genetic-engineering techniques
    c. Next generation robotics
    d. Neuromorphic technology
    e. Distributed manufacturing

12.12 Summary

- We discussed big data characteristics, technologies and tools in detail.
- We presented under-pinning technologies for the evolution of big data and discussed an emerging big data computing platform.
- We also discussed open-source technologies that are emerging big data visualization today.
- The big data scientists are thinking outside the box and using ensemble models to generate the best predictive results.
- Big data technologies are enabling faster and pulsing to turn the data into insights.
- Big data scientists prefer get scrappy to solve real-world problems with big data.

12.13 Glossary

Visualizing Data: Visualizing data is a technique to identify patterns in the data and processing the data to make it more consumable.

Data Artisans: The people who create these new visualizations are called 'data artisans'. These data artisans are skilled in science, design, and art.

Sample: A sample consists of one or more observations from the population.

Population (all the data): A population includes all of the elements from a set of data.

Get Scrappy: Being Get scrappy means committing to a result at all costs, and doing whatever it takes to get that result.

Emerging Technology: Emerging technologies are those technical innovations which represent progressive developments within a field for competitive advantage.

Ensemble Models: Ensemble modeling is the process of running two or more related but different analytical models, and then synthesizing the results into a single score or spread, IS in order to improve the accuracy of predictive analytics and data mining applications.

12.14 Self-Assessment Test

1. What is visualization? How it is making big data more consumable?
2. How is visualization enabling organizations to take immediate decisions?
3. List some open-source project names.
4. What is the difference between Sample and Population?
5. Why are big data analytics moving from sampling to using all the data?
6. List some open-source emerging technologies that support big data analytics.

12.15 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition, April 2021.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

12.16 Answers to Check Your Progress Questions

1. (d) Discovering

Business analytics use graphs, charts, and dashboards to represent the data into comprehensive format. The intent ‘Discovering’ interacts with the data to understand relationship in data.

2. (c) Artisans

The people who create these new visualizations are called ‘Data Artisans’. These data artisans are skilled in science, design, and art.

3. (d) Data visualization

The data visualization is making big data analytics iterative. It is also reducing cycle time of big data analytics so that immediate action can be taken.

4. (d) **More or all of the data**

Big data scientists are using more or all of the data instead of a sample so that it increases the accuracy of the model.

5. (c) **Ensemble models**

Big data scientists are generating better results for their enterprise. They are using multiple techniques to create models to meet business objectives. By using multiple techniques, they are creating multiple models which are called 'ensemble models'.

6. (c) **Re-evaluated**

The models deployed periodically and reevaluated by the creators is called ad 360 degrees modeling. The frequency of reevaluation is very important and varies tremendously.

7. (d) **Fast data**

Big data analytics and its insights will always be a part of forecasting, but fast data will enable businesses to make decisions in real time as it flows into the organization.

8. (a) **Get scrappy**

Get Scrappy enables you with the courage and commitment to pursue your goals even though there is no evidence that you can succeed.

9. (c) **R**

Many software players are working in an open-source environment towards big data analytics. R is an open source statistical tool that explores and builds models.

10. (d) **Neuromorphic technology**

These emerging techniques are designed to solve specific business problems which incorporate deeper and more complex prescriptive analytics. Neuromorphic technology makes computer chips that mimic the human brain.

Big Data, Cloud and Analytics

Course Structure

Block 1: Introduction and Applications of Big Data	
Unit 1	What is Big Data?
Unit 2	Why Big Data is Important?
Unit 3	Big Data in Marketing & Advertising
Unit 4	Big Data in Healthcare
Block 2: Cloud Computing and Big Data Technologies	
Unit 5	Big Data and Cloud Technologies
Unit 6	Big Data Technologies and Terminologies
Unit 7	Cloud Computing and Big Data Management for Decision Making
Unit 8	Handling Unstructured Data
Unit 9	Information Management
Block 3: Business Analytics	
Unit 10	Analytics in Database Marketing
Unit 11	Business Analytics Techniques
Unit 12	Data Visualization and Modelling
Block 4: Managing Talent for Big Data Analytics	
Unit 13	Talent Management-I
Unit 14	Talent Management-II
Block 5: Data Privacy and Analytics in Various Business Areas	
Unit 15	HR Analytics in HR Planning
Unit 16	Data Analytics for Top Management Decision Making
Unit 17	Business and Marketing Intelligence Using Analytics
Unit 18	Data Privacy and Ethics