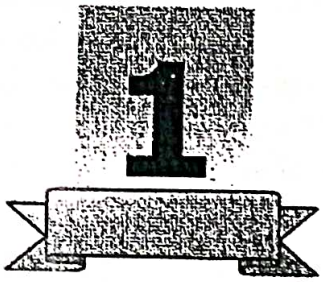


Introduction : Data Science and Big Data

Syllabus

At the end of this unit, you should be able to understand and comprehend the following syllabus topics :

- Introduction to Data science and Big Data
  - Defining Data science and Big Data
  - Big Data examples
  - Data Explosion
    - Data Volume
    - Data Variety
    - Data Velocity
    - Data Veracity
  - Big data infrastructure and challenges
- Big Data Processing Architectures
  - Data Warehouse
    - Re-Engineering the Data Warehouse
    - Shared everything and shared nothing architecture
- Big Data Analytics
- Data Science - The Big Picture
- Machine Learning
  - Big data learning approaches
  - Relation between AI
  - Statistical Learning
  - Data Mining

1.1 Introduction to Data Science and Big Data

Q. What is Big Data ? **SPPU - Aug. 18 2 Marks, May 19, 3 Marks**

- Increasingly people and things are getting interconnected. Data is continuously created by devices and users. For example, when you go for online shopping all your clicks and views, your interaction with the website, your interaction with the competitor website, your addition to the cart, removal from the cart, price comparison, review reading, etc. are all recorded to analyse you as a user and a prospective buyer who could be influenced to make a purchase. The entire agenda of conducting data analytics is based on making informed decisions that can be further used to shape your behaviour and drive the business intentions.

- Have you heard about the company Cambridge Analytica? It was a political consulting firm that harvested data of about 87 million US voters during Trump's presidency campaign in 2014. It built a system that could profile individual US voters in order to target them with personalised political advertisements. The result, everyone knows!
- Data analytics combined with the right set of data is an extremely powerful mechanism today for businesses and nations. It can be used to derive meaningful predictions and shape user behaviour.
- The term Big Data (always write capitalized B and D in Big Data. Big Data is a noun that has a special meaning!) refers to an accumulation of data that is too large and complex for processing by traditional database management tools. The datasets referred under Big Data are massive and could be petabytes in size.

Definition : *Big Data consists of extensive datasets that require a scalable architecture for efficient storage, manipulation, and analysis.*

- Big Data problems usually require specific set of tools and techniques for processing and usually are not traditional database systems. The data to be processed could be historical or could be accumulated in real time.

Definition : *Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.*

- Data science as a concept unifies several domains such as statistics, data analysis, informatics, and their related methods in order to understand and analyse real-world phenomena with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights on business data. Data science is heavy on computer science and mathematics. Data science is used in business functions such as strategy formation, decision making and operational processes. It touches on practices such as artificial intelligence, analytics, predictive analytics and algorithm design.
- Data science is related to data mining, machine learning and big data.

1.1.1 The Five Vs (Characteristics) of Big Data (Data Explosion)

Q. Explain characteristics of Big Data. **SPPU - Aug. 18, 2 Marks**

Q. Explain 3 V's of Big Data. **SPPU - May 19, 5 Marks**

Big Data is often characterised as 5 Vs. Fig. 1.1.1 shows the characteristics of big data.

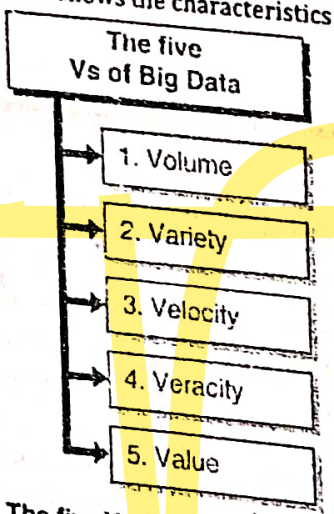


Fig. 1.1.1 : The five Vs (characteristics) of Big Data

1. Volume

- Volume refers to the size of the dataset. It could consists of billions of rows and millions of columns. It typically ranges from terabytes to petabytes of data.
- For example, imagine a dataset of all active users on flipkart.com. It could be a listing of all users and the clicks they made in a day. Assuming a user showed interest in 5 different mobile handsets and clicked on 10 photos each it would mean 50 data points per user. Now, if flipkart.com has 1 million users looking at mobile handsets on a sale day, it could mean 50 million data points!
- Usually such datasets are stored in multi-tiered storage media. The data storage and retrieval could be time consuming.

2. Variety

- Variety refers to the data accumulated from multiple data sources.
- There could be various types of data collected in various formats and structures. The data could be structured, semi-structured or unstructured.
- For example, data on cricket world cup could mean scoring tables, videos, audio, tweets, texts, images, comments, and news reports. Such a wide variety of data on a particular topic could make Big Data analysis further complex.
- The data is carefully chosen so as to be meaningful for the purpose of analysis and objectives to be met.

3. Velocity

- Velocity refers to the rate of speed at which the new data is generated and processed. It could be generated in real time or could be historical in nature.
- For example, in a live match, several new tweets could be generated in a second. If you were to write a Big Data application that analyses tweets in real time, it could mean ingesting tweet data at the rate of millions per second.
- Many Big Data applications process real time data such as "likes" and "shares" to report trends and top of the hour news. But, keep in mind that Velocity may not always be high. For example, medical records could have been compiled over several years before they are analysed.

4. Veracity

- Veracity refers to the integrity of the Big Data. It is a measure of data quality and usefulness of the data.
- For example, a dataset on cancer patients could have low veracity if it has data for non-cancer patients as well. Such a data could produce inaccurate insights and may not be useful for carrying out Big Data analysis. You should ensure that the data chosen for Big Data analysis is accurate and is free from biases, noises, and abnormality.

5. Value

- Value measures the degree of usefulness of data for the organisation.
- The longer it takes for the data to generate meaningful results, the lower the value of the data. In a way, Value of data is dependent on the Veracity of data. High Veracity data generally has high value for the organisation.
- Value of data also depends upon
  - (a) How it was stored ?
  - (b) How old / recent it is ?
  - (c) How the data attributes were preserved ?
  - (d) What questions can be answered with the data ?

1.1.2 Major Applications of Big Data

Today, Big Data analysis is carried out for various purposes. However, some of the most common application of Big Data analysis are shown in Fig. 1.1.2.

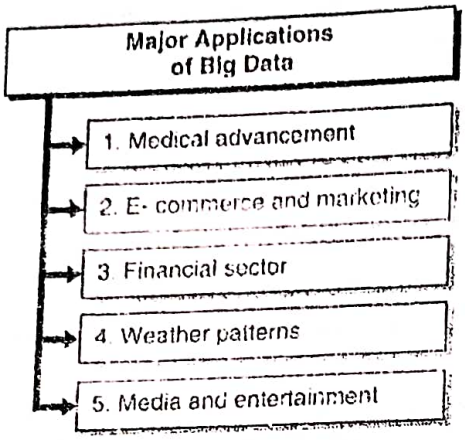


Fig. 1.1.2 : Major Applications of Big Data

1. Medical advancement

- The healthcare industry is one of the biggest proponents of Big Data analysis.
- Large data sets are used for research and operations in various areas such as :
  - (a) Life threatening diseases such as cancer and AIDS
  - (b) Reduce healthcare costs
  - (c) Find cure for new epidemics and viruses
  - (d) Predict the diseases, such as Alzheimer's, in early stages to improve chances of survival
  - (e) Genetic sequencing to predict genes based diseases
- Big Data analysis provides a way to look at various patient's data and build models around which new study and research can be carried out.

2. E-Commerce and Marketing

- Identifying the next product or marketing effort that would yield results is a great way for companies to build adequate strategies.
- Consumer patterns, their likes, their purchase potential, time of the year when the users buy, market trends, global economy etc. are all accounted to come out with the new products and services.
- The consumer's data is analysed to identify patterns and make predictions about the future demands.
- For example, if you are shopping early stage pregnancy products, it is likely that you would purchase infant's products soon. Companies can use this information to push new schemes and offers to you and also equip their warehouses to ship baby products faster to you by placing the inventory closer to you.

3. Financial Sector

- Big Data analysis is also heavily used in the financial sector.
- Some of the common applications of Big Data analysis in financial sector are
  - (a) Detecting frauds. For example, if you made a physical transaction in Delhi and then another physical transaction in Mumbai within one hour, it is highly unlikely that you could have made both the transactions.
  - (b) Credit score analysis based on your past reputation to pay dues
  - (c) Quoting customised insurance premiums based on your life style
  - (d) Selling new products and services to you. For example, if you have surplus balance in your account throughout the year, then you could be sold investment products.
- Apart from these, the financial sector also uses Big Data for detecting money laundering, shell companies, fraudulent transactions, and reporting. Based on the transactions carried out by individuals, the companies can build financial health profile of its consumers and identify future spend patterns and requirements.

4. Weather Patterns

- Big Data analysis is crucial for detecting changes in the weather patterns. You would have heard about
  - (a) The rising ocean temperatures
  - (b) Global warming
  - (c) Melting glaciers in Antarctica
  - (d) Reducing Oxygen level
- There is a huge amount of data that can be used to predict the weather changes and report how it is affecting our environment. It can be used to predict weather forecast, natural disasters and any other changes that could affect our well-being.

5. Media and Entertainment

- The media and entertainment industry use Big Data analysis to understand viewing and liking patterns for the media content.
- Based on the time of the day, season, device you are on, your personal interests and taste, the content can automatically be recommended for you. I am sure you would have seen YouTube recommendations as you watch YouTube videos. Similarly, companies like Spotify can automatically create curated and customised playlists for you based on your listening profile.

1.1.3 Data Formats

Big Data comes in different formats. Data can be machine generated (such as log files) or could be human generated (such as tabular data). Overall, the data format is classified as shown in Fig. 1.1.3.

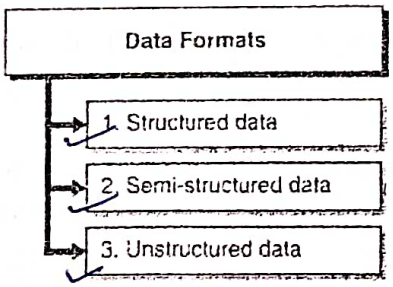


Fig. 1.1.3 : Data Formats

1. Structured Data

- Structured data exhibits a particular order (also known as model or schema) for storing and working with the data. The data attributes are usually related and are often the basis of analysis.
- The structured data is usually generated by machines or compiled by humans.
- For example, spreadsheets, customer records, transaction records, sales reports, etc. are all structured data. The structured data is usually stored in relational databases or simple CSV or spreadsheet files.

2. Semi-structured Data

- Semi-structured data has some definitive patterns for storage, but the data attributes may not be inter-related.
- The data could be hierarchical or graph-based in nature. The semi-structured data is usually stored in text files as XMLs, JSON or YAML format. The common sources for semi-structured data is usually machines such as sensors, website feeds, or other application programs.

3. Unstructured Data

- Unstructured data does not exhibit a fixed pattern or a particular schema. This is the most common format of Big Data.
- Examples of unstructured data are video, audio, tweets, likes, shares, text documents, PDFs, and scanned images. Special tools and mechanisms are required to process unstructured data. Also, it is usually cleaned (sanitised) before it can be used for analysis.

1.1.4 Comparison between Data Formats

Table 1.1.1 : Comparison between Data Formats

Comparison Attributes	Structured Data	Semi-structured Data	Unstructured Data
Volume of Data	Low	Medium	High
Processing Complexity	Low	Medium	High
Data generated by	Humans and Machines	Machines	Humans
Data usually stored in	Relational Databases	Textual files	Binary files
Patterns and Schema	Fixed	Flexible	Random
Specialised Tools	Not required	Not required	Required

1.1.5 DIKW Pyramid

- It is important to understand how data can be enriched and the journey it takes with each stage of enrichment. To understand this journey, typically DIKW Pyramid is referenced. DIKW is an acronym for the four stages of data enrichment that are

1. Data
2. Information
3. Knowledge and
4. Wisdom

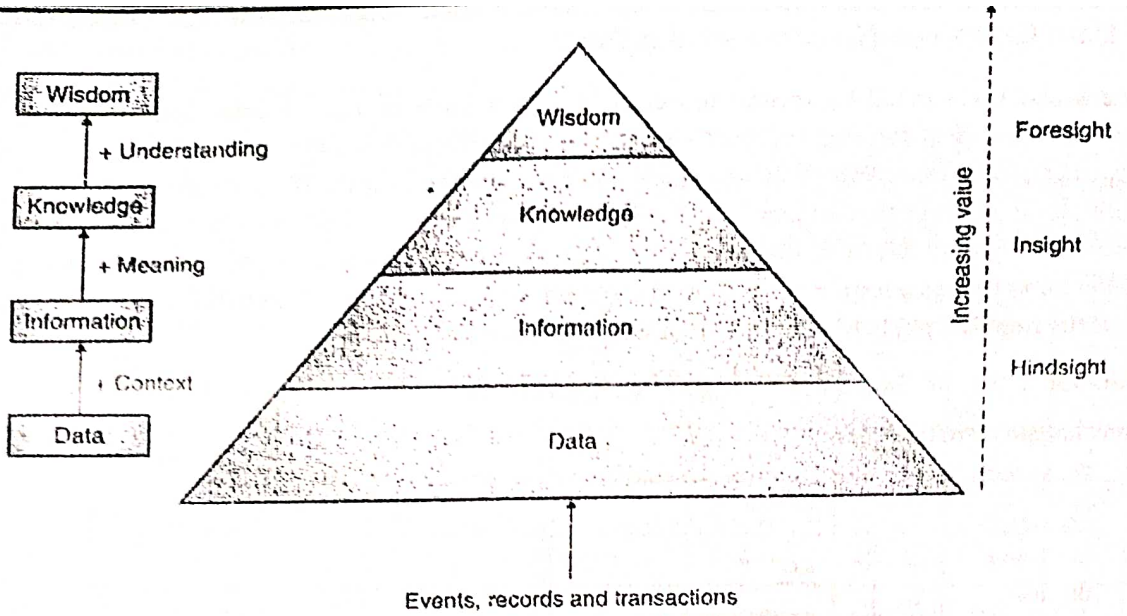


Fig. 1.1.4 : DIKW Pyramid

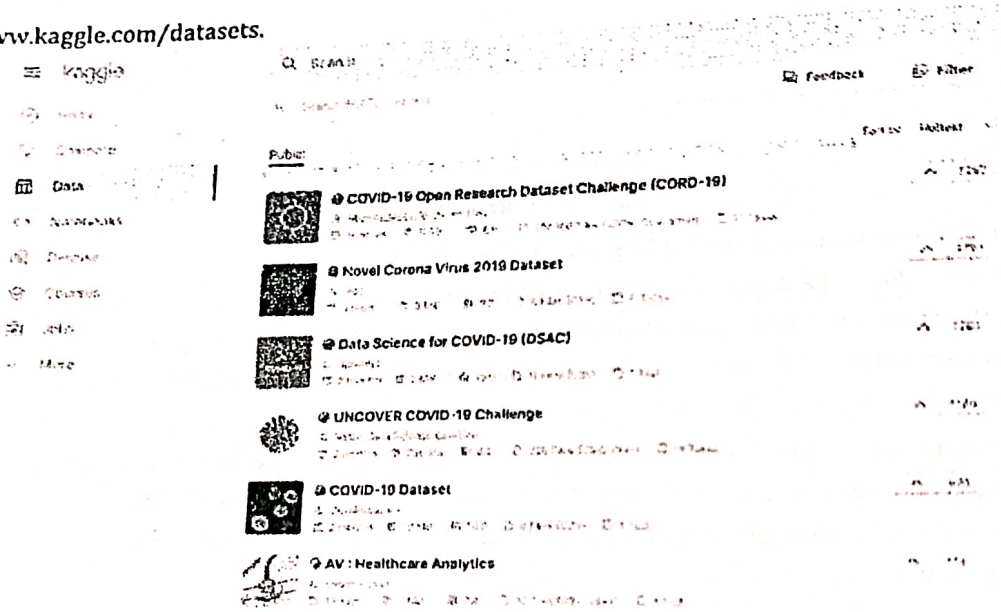
1. **Data :** This is the lowest bottom of the DIKW pyramid. This is the raw data collected from various events, records, and transactions around you. The data could be generated by machines or by humans. The data itself does not have very high value until it is enriched with more attributes that can be used for further analysis. For example, you could just have a data set that has list of millions of people with their demographic details and how they died. This data may not give you anything actionable.
2. **Information :** At the next level, the collected data is enriched with the context to give information. You start to build perception about the data to give you hindsight. The hindsight about the information reflects or acknowledges what is contained in the data. For example, you could begin to see that the list of people is actually the list of cancer patients and their life patterns.
3. **Knowledge :** When you add meaning to the information, you start to gain knowledge. This is where you precisely start analysing the information at hand and make it more useful and meaningful. You could gain deep insights about the information and be able to answer high-level questions. For example, you can analyse the information on cancer patients and build patterns around life expectancy after cancer detection with or without chemotherapy. You could further analyse the effect of various chemotherapy medicines to understand their dosage and effectiveness level. A company could then invest in building more effective chemotherapy medicines to improve life expectancy of cancer patients after diagnosis. So, understand here that the objectives of data analysis must be clear to derive meaningful knowledge from the information at hand.
4. **Wisdom :** The final level of wisdom is achieved when you add understanding to the derived knowledge. Note here that wisdom is not achieved using a technical algorithm or formula but is based on the human understanding of the data analysis that was carried out. For example, after analysis of data on cancer patients, you could understand what lifestyle to follow in terms of diet, sleep, and exercise to avoid or delay occurrence of cancer. This understanding could be shared with the world as foresight.

As you see, the DIKW pyramid helps to put a perspective on visualising the data analytics stages. Now let's take this discussion further and understand the categories of data analytics.

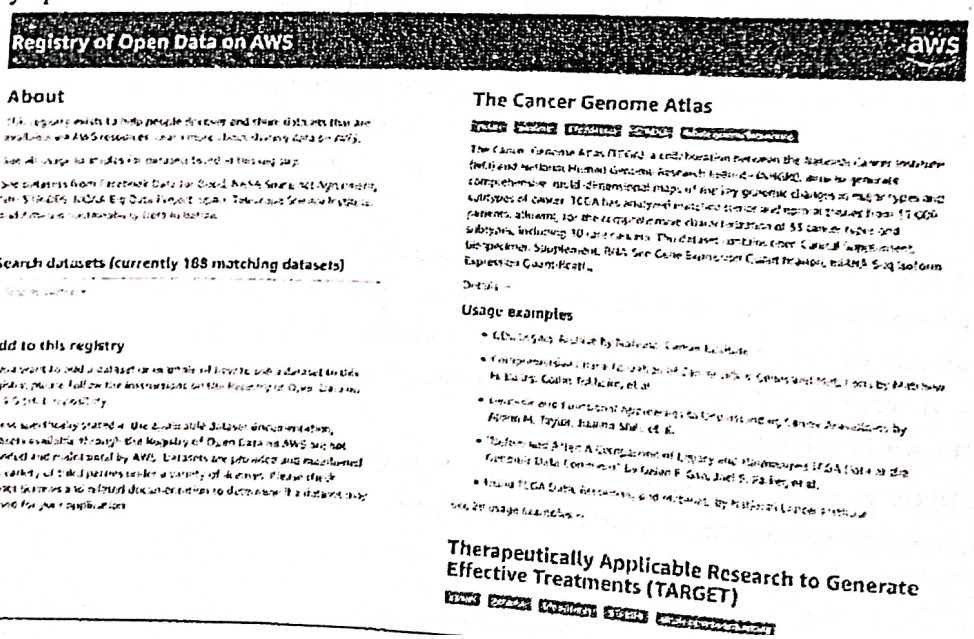
1.1.6 Big Data Examples (Sources of Big Data)

- As you understand, there could be several sources of Big Data such as data collected from machines, weather data, social media posts, surveys, tweets, and what not. As you understand, to build a Big Data Analytics model, you need a huge volume of training data. You could possibly have this training dataset available internally from your historical business operations or you could engage with external agencies and websites that have training datasets for various purposes either freely available or available for a fee. Wherever you get data from, you need to ensure that it is accurate and verified to be as real as possible. Your trained Big Data Analytics model would only be as accurate as your training data. Some of the popular, publicly available, dataset sources are as following.

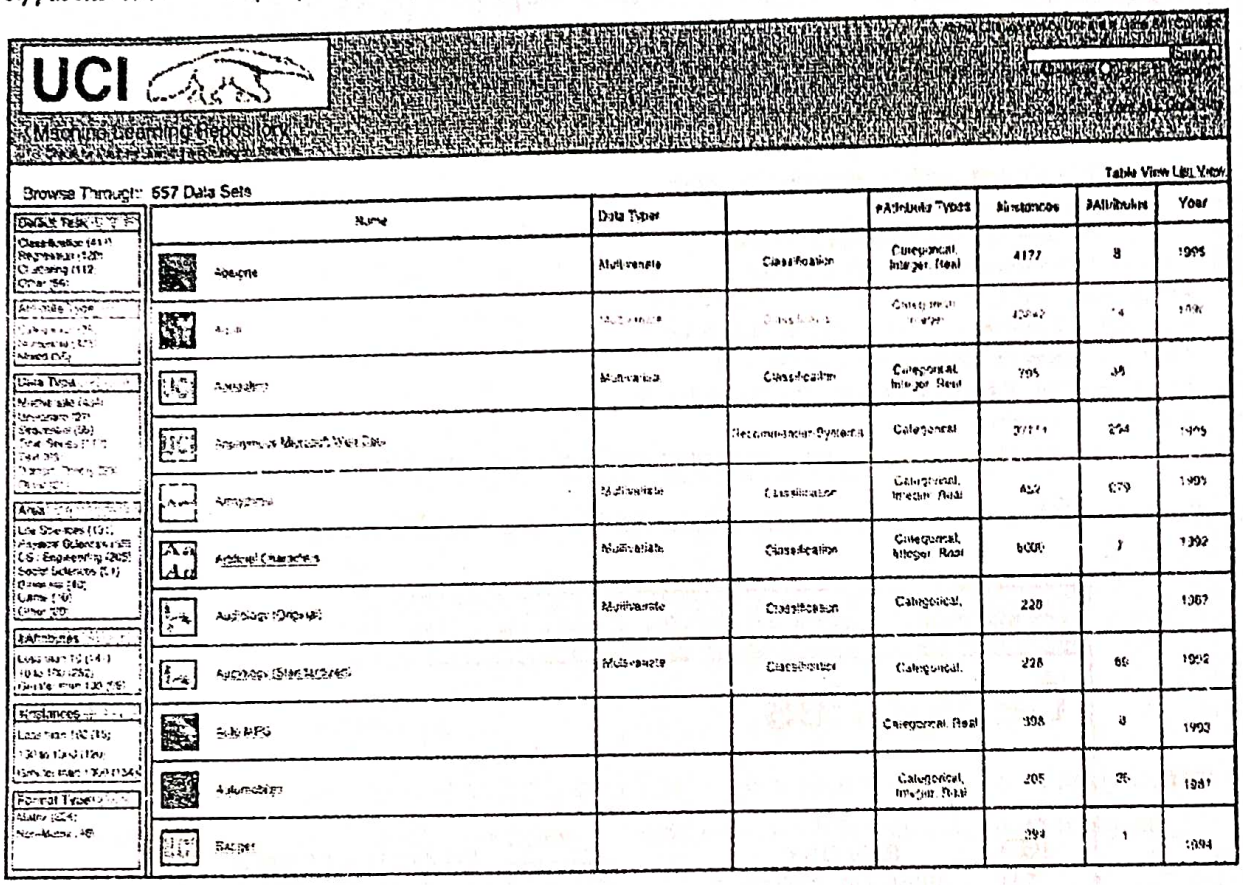
a. **Kaggle Dataset** : As of Sep 2020, Kaggle has 54,876 datasets on various topics. They are available on https://www.kaggle.com/datasets.



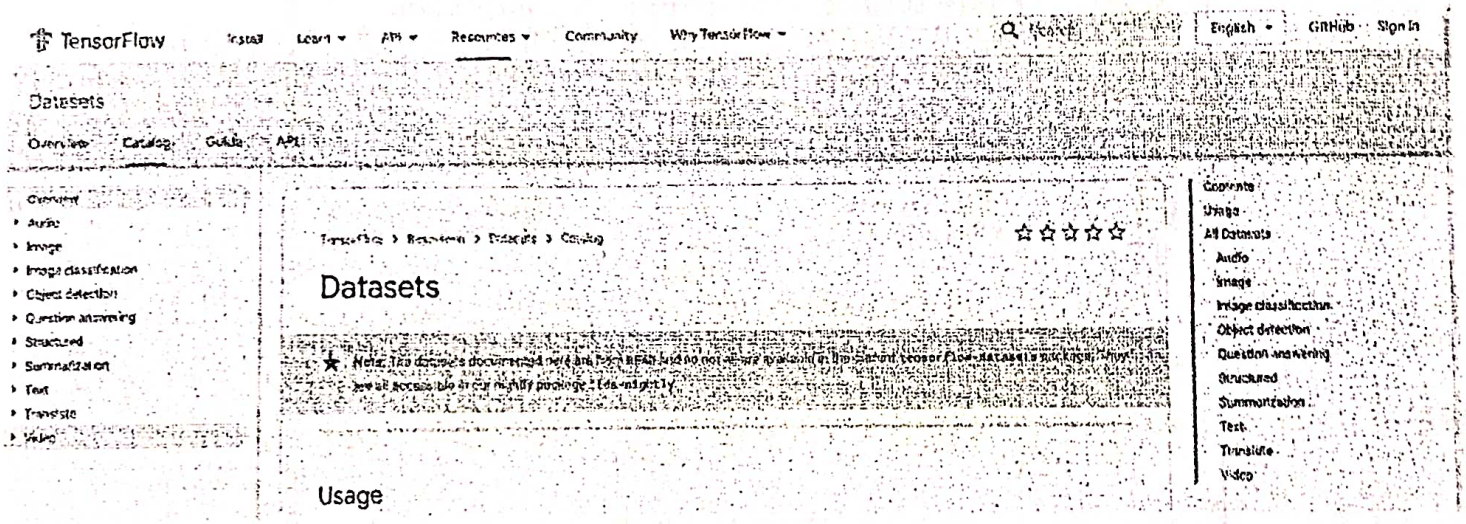
b. **Amazon Web Services (AWS)** : As of Sep 2020, AWS has 188 datasets on various topics. They are available on https://registry.opendata.aws.



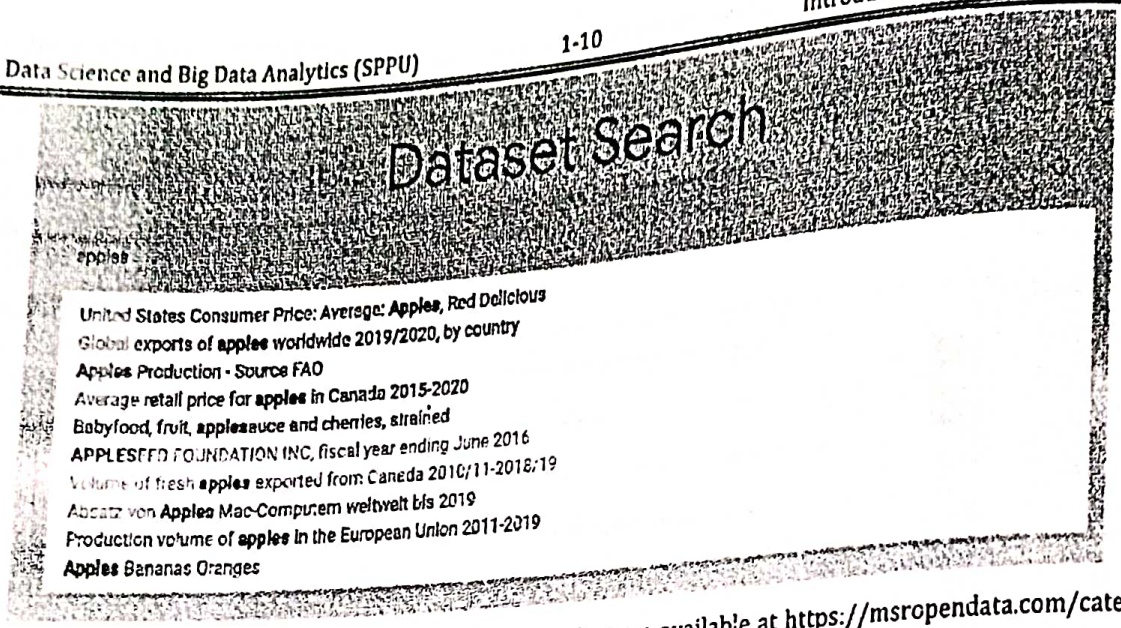
c. **UCI Machine Learning Repository :** As of Sep 2020, UCI has 557 datasets on various topics. They are available on https://archive.ics.uci.edu/ml/datasets.php.



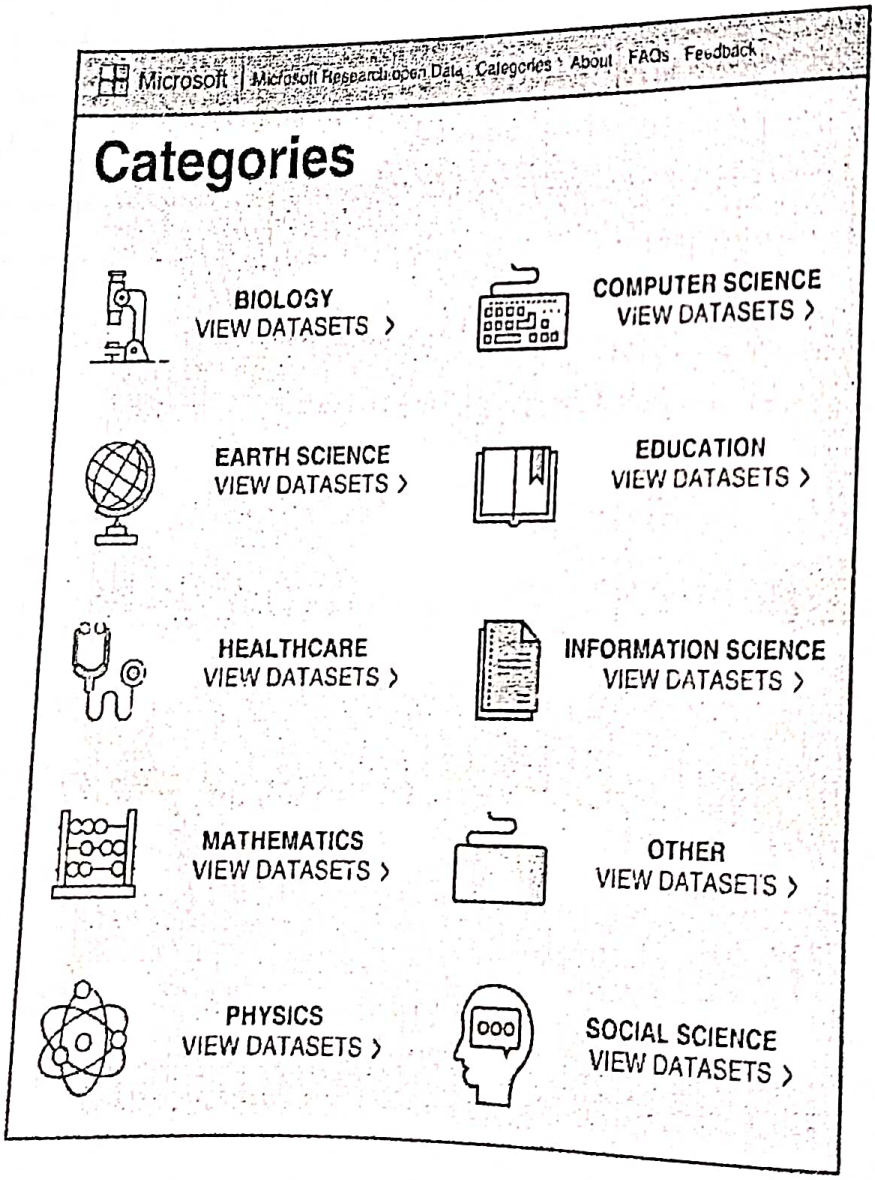
d. **Google's TensorFlow :**Google has a huge repository of various datasets – audio, video, images, text, and various others. They are available at https://www.tensorflow.org/datasets/catalog/overview.



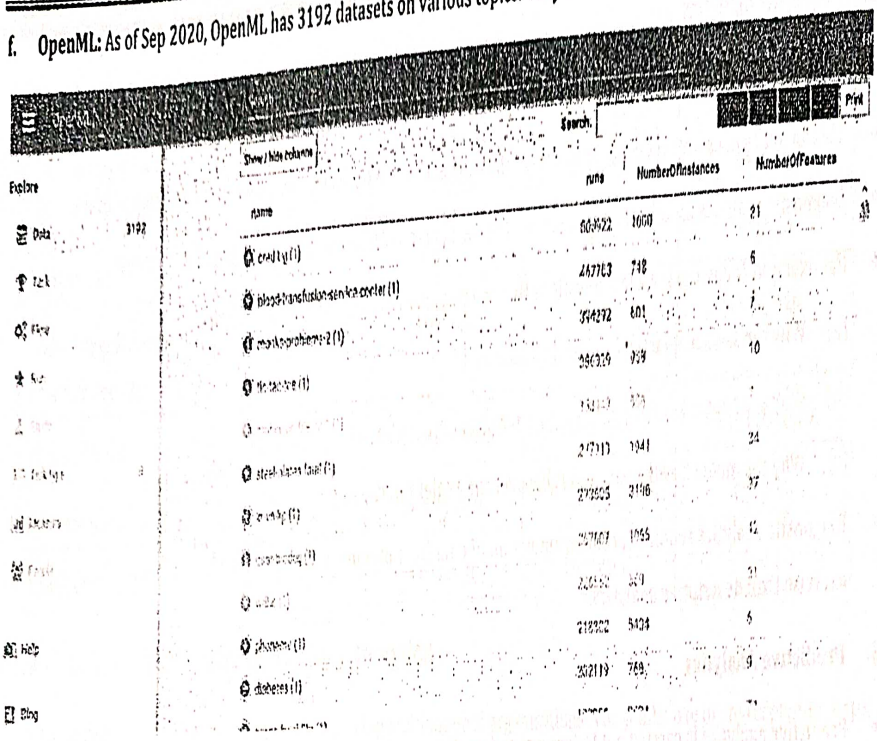
Google also provides a search option for finding various datasets across various dataset repositories at https://datasetsearch.research.google.com.



e. **Microsoft :**Microsoft has a repository of various datasets. They are available at https://msropendata.com/categories.



f. **OpenML:** As of Sep 2020, OpenML has 3192 datasets on various topics. They are available on https://www.openml.org.



So, as you see, there are quite a many dataset repositories around from where you can collect data.

1.1.7 Categories of Data Analytics

Now that you have a fair understanding of what Big Data is, let's touch upon the spectrum of analytics that is possible on the given data sets. Fig. 1.1.5 shows the different types of analytics require different tools and techniques.

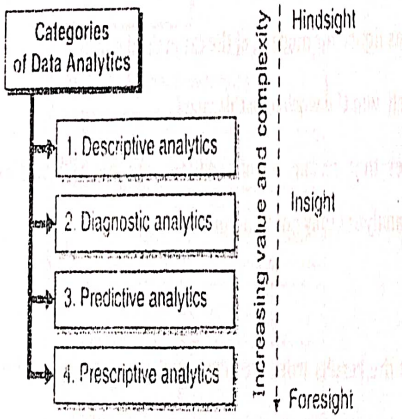


Fig. 1.1.5 : Categories of Data Analytics

1. **Descriptive Analytics**

- Descriptive analytics answers questions about events that have already occurred. Usually raw data is queried without adding any contextual information.
- This is the simplest form of analytics and typically answers questions such as
  (a) How many units of a particular item sold in last 6 months?
  (b) How many patients died of a particular cancer type?
  (c) How many calls did you receive for a particular issue?
- This kind of analytics is usually done using database queries or simple spreadsheet filters. You could have periodic dashboards and reports that can be used to visualise results of the descriptive analytics.

2. Diagnostic Analytics

- Diagnostic Analytics is done to find out cause of a phenomenon or derive reasoning behind events.
- This analytics goes a level deeper to provide information that can be used to fix a particular situation or event.
- Diagnostic analytics usually adds more context to the data to get information about a particular interest.
- For example, following are a few questions that can be answered using diagnostic analytics.
  (a) Why the sales in quarter 2 lower than quarter 1?
  (b) Why are people falling ill after eating a particular type of biscuits?
  (c) Why the model X of the car preferable over the model Y of the car?
- Diagnostic analytics require careful examination of data from multiple sources and is a little more involved and skilful exercise than descriptive analytics.

3. Predictive Analytics

- Predictive analytics is carried out to forecast and predict future events.
- The information is further enriched by adding meaning to it to derive knowledge.
- The predictive data models are carefully created that can base off future predictions based on the past events. Predictive analytics could possibly answer questions such as
  (a) What would be the improved life expectancy if choosing medicine A over medicine B?
  (b) What would be the sales figure for model X of the car in third quarter?
  (c) Which team would likely win the world cup this year?
- Predictive analytics assumes that certain set of conditions are met or would exist. If there are changes to those conditions, then predictive analytics may not be accurate.

4. Prescriptive Analytics

- Prescriptive analytics takes the results from predictive analytics and further adds human judgement to prescribe or advise further actions.
- This reflects the wisdom level from the DIKW pyramid that you learnt earlier.
- The prescriptive analytics could answer questions such as
  (a) What should you do to delay cancer?
  (b) What is the best time to leave home to reach airport on time?
  (c) Which medicine would have higher chances of survival for the patient?
- Prescriptive analytics is the most difficult out of all other analytics. It requires significant skills and time to give effective actions and results. It could also be dependent on not only the analysed data but external conditions such as political pressure, social acceptability, and personal preferences.

1.1.8 Comparison between Categories of Data Analytics

Table 1.1.2 : Comparison between categories of Data Analytics

Comparison Attribute	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Complexity	Least	Medium	High	Highest
Time requirement to produce results	Low	Medium	High	Very High
Value of results	Short Term	Medium Term	Long Term	Very long term
Data enrichment level	Data	Information	Knowledge	Wisdom
Analytics Frequency	Most Common	Frequent	Not often	Rare

1.1.9 Drivers (Motivation) of Big Data Analytics

- A few decades back, computing was restricted to big multinational organisations. The public use of smart phones, apps, internet connectivity was negligible if not absent.
- Businesses and Governments were not so agile and the use of data for business growth and optimisation was negligible. But, the increased use of computers with internet connectivity has changed practices around data analytics.
- Fig. 1.1.6 shows some of the major drivers or motivations for Big Data analytics :

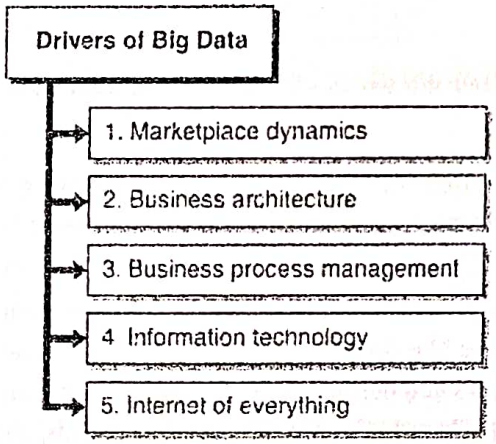


Fig. 1.1.6 : Major drivers or motivations for Big Data analytics

1. **Marketplace Dynamics**

Today we live in a global economy. Recession in one country's economy affects the other.

Companies are increasingly looking to optimise their operations and reduce costs and at the same time acquire new customers and retain the existing ones. The market demand in a global economy can be uncertain and hence companies need to invest in data analytics to understand trends and forecast business decisions.

The historical perspective from business intelligence reports alone is no more sufficient. For example, there was a recession in 2008. Businesses today are already trying to predict the next recession or economic slowdown and appropriately cut workforces or streamline their business ahead of time to minimise the effect of next recession.

2. Business Architecture

Business architecture is a means of articulating business design and how it operates. Almost all sizable and meaningful businesses today are utilising the power of IT to transform the business and make it more agile.

The business architecture links the broad perspectives such as business mission and vision to business services, organisation structure and key performance parameters to gauge the level of success.

Big Data ties such performance parameters to organisation's layers right from the bottom most layer to the executive level. At each layer, it can precisely monitor key performance parameters and help business to make actionable and right decisions.

For example, it can establish that the production unit must make 2,400 items to meet the festival season demands. Now, to meet that level of production what all needs to happen is something that the business architecture can decide.

3. Business Process Management

Business processes define how work is performed in an organisation.

Business process management aims to continuously improve the business processes and it uses Big Data as input to understand the improvement opportunities.

For example, if 90% of the diabetic patients require blood and urine check up on every visit, it might make sense to move the check-up unit closer to the diabetic patients' waiting area.

Patients do not need to walk for long within the hospital to give samples and then later collect reports. Such data driven business process decisions have become key to providing a great user experience.

4. Information Technology

Information technology has connected the world at hyperscale. Almost half of the planet's population has internet connectivity.

There is a huge volume of user data from social media apps, audio and video streaming media services, hospitals and laboratories around the world, government surveillance of public areas using CCTV cameras, crime intelligence reports, biometric information such as face patterns and fingerprints and plenty of other sources.

The vast digitisation of everything has made it easy to collect the data and use it for rendering value added services and enhanced user experience. Companies like Google and Facebook know almost all your personal preferences such as places you visit, food you like, activities you perform, friends you hang out with and several of your other demographic details. You are then targeted for specific products and services through ads or other marketing campaigns.

Also, computing in general has become cheaper with commodity hardware and availability of cloud computing. Any organisation can now process the vast amount of user data in almost real time and derive meaningful insights. Advancements in information technology has truly made an impact on Big Data analytics by connecting users to make more and more data available for analytics and at the same time making it comparatively easier and cheaper to process the collected data.

5. Internet of Everything

It is no more just the people who are getting connected to the internet. The day to day use things around you are getting connected as well. Smart refrigerator, microwave, lights, AC, geyser, toaster, CCTV cameras, music player, TV, watch, cars and perhaps what not.

Each of these devices collect huge amount of data and passes it on the companies who can use them to improve existing business services and provide newer ones. For example, your smart watch can automatically monitor your heart rate patterns and book an appointment with a cardiologist when it deems that the situation requires medical consultation and advice. To establish such predictive models, huge amount of data is collected from various users and the event of an abnormal heart rate pattern to a cardiologist visit is correlated.

1.1.10 Emerging Big Data Ecosystem and New Approach

Q.	Explain Big data ecosystem.	SPPU - Aug. 18, 4 Marks
Q.	Discuss with example Data devices and Data collectors of emerging Big Data Ecosystem.	SPPU - Oct. 19, 5 Marks
Q.	Draw and Explain Big Data Ecosystem.	SPPU - Dec. 19, 5 Marks

- In the emerging Big Data ecosystem, there are four main groups that work together seamlessly to make the most out of the Big Data analytics. They are as shown in Fig. 1.1.7.

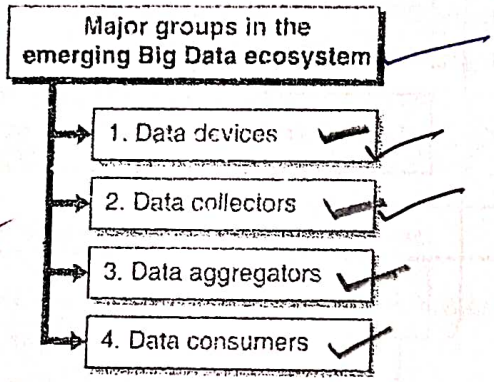


Fig. 1.1.7 : Emerging Big Data Ecosystem

1. **Data Devices**

Data devices are the physical equipment that collect the data. There can be a wide variety of data devices such as smart phones, sensors, CCTV cameras, computers, and other digital and connected equipment. Based on their form factor and purpose they can collect various types of data and send it to the desired target for data analytics.

2. **Data Collectors**

Data collectors are institutions or organisations that facilitate collection of the data. For example, it could be an online ecommerce portal or a local retail store. Which products you look at these stores, what you buy, your preferences, your buying habits, etc. are all known and collected as you shop through. Similarly, the audio and video streaming services record your preferences and viewing patterns.

3. **Data Aggregators**

These are larger institutions or organisations that collect and accumulate the data from various data collectors. For example, VISA could collect the credit card and debit card related data irrespective of which bank's card you use. Similarly, Google can collect your location data irrespective of which city or country you go to. Note here that data collector and data aggregator could be the same company as well. For example, Google could collect data from individual Google services that you use such as Google Maps and Gmail but also aggregate the data across all its services.

4. Data Consumers

Finally, Data Consumers are institutions or organisations that buy or use the collected and aggregated data. These are the organisations that ultimately make use of the data and benefit from its analysis. For example, a bank could purchase the data of individuals to find out who could take a home loan and then approach such individuals. The ecommerce portal could get information about the browsing patterns of the users and could suggest the recommended products automatically when the particular user visits its own website.

Again note here that a company could be a data collector and a data aggregator and also a data consumer. It really depends on the size of the company, its presence across the user ecosystem and its capability to collect, aggregate, analyse and use the data.

The Fig. 1.1.8 shows a high-level relationship between data devices, collectors, aggregators, and consumers.

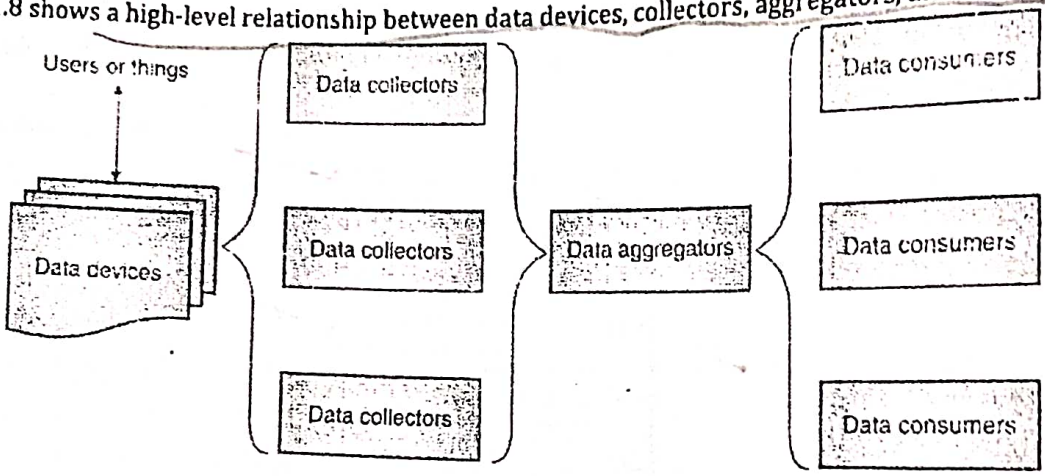


Fig. 1.1.8 : High-level relationship between data devices, collectors, aggregators, and consumers

Taking the examples in the case of Cambridge Analytica company for its role in US Presidential elections in 2014,

- Users used their smartphones or computers to access Facebook. Such devices are "Data Devices" in the Big Data ecosystem.
- Facebook had the use profile and other information such as timelines and friends. Hence, it is "Data Collector" in the Big Data ecosystem.
- Cambridge Analytica aggregated the data of millions of Facebook users. Hence, it is "Data Aggregator" in the Big Data ecosystem.
- Finally, Cambridge Analytica sold the user's personal information to the political parties for targeting their ad campaigns. Hence, political parties are "Data Consumers" in the Big Data ecosystem.

1.1.11 Key Roles in the New Big Data Ecosystem

Fig. 1.1.9 the key roles in the new Big Data ecosystem.

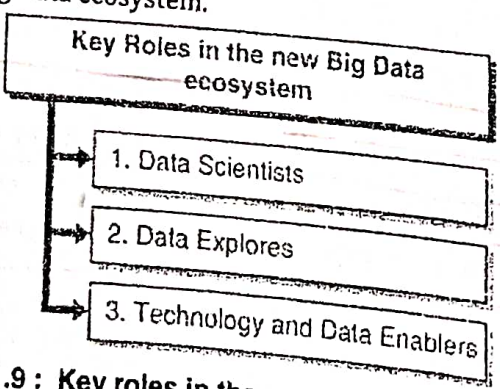


Fig. 1.1.9 : Key roles in the new Big Data ecosyst

1. **Data Scientists**

These are individuals with deep analytical talent required to understand and work with data. They build sophisticated data models that can make predictions or carry out the required analytics. They have advanced training to understand and work with data. They have expertise and vast background in statistics, mathematics, and economics.

2. **Data Explorers**

This group of individuals has less analytical depth than data scientists. Their primary role is to understand the data required for answering the questions at hand and ensuring that the right data is collected, cleaned, and loaded for analysis. The individuals typically fitting this role are financial analysts, business analysts, market analysts, operations managers, sales managers, etc.

3. **Technology and Data Enablers**

This group of individuals is more technical oriented, and they understand how to handle and program for data analytics at scale. They have broad understanding of the software tools, hardware requirements, storing Big Data and have other technical expertise required for carrying out data analytics. Typical roles fitting this category are data architects, solution architects, system engineers, Big Data programmers, etc.

1.1.12 Key Roles for a Successful Analytics Project

Q. Enlist and explain various users involved to make successful analytical project. **SPPU - Oct. 19, 5 Marks**

A successful data analytics project requires several people. Fig. 1.1.10 shows key roles for a successful analytics project.

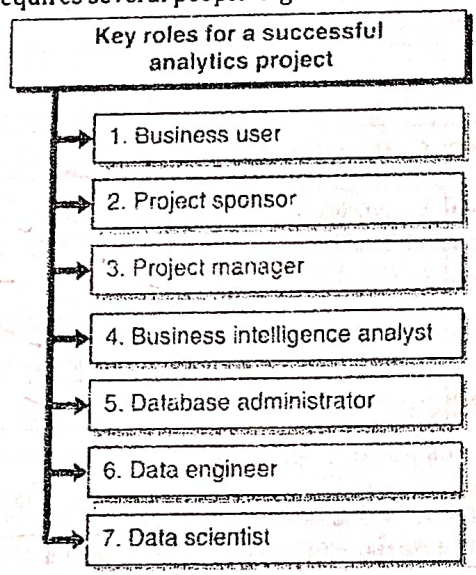


Fig. 1.1.10 : Key roles for a successful analytics project.

1. **Business User :** Business user is an individual who uses the results of data analytics to meet the business objectives. This is usually a business analyst, an operations manager, or a subject matter expert.
2. **Project Sponsor :** This is usually an executive or a senior management staff that provides approval and funding for the project. She sets the business problems to solve. She monitors the health of the project and ensures that it is progressing towards established goals and desired business outputs.
3. **Project Manager :** Project manager is responsible for day to day execution of the project. She ensures that the project milestones are achieved, and the overall project is running timely.

4. **Business Intelligence Analyst :** Business Intelligence Analysts provide business domain expertise. They understand business and its key performance parameters and metrics.
5. **Database Administrator :** Database administrators setup, operate and maintain the databases that hold the actual data for analytics and the results of analytics. They are responsible for ensuring that the database is up and running and only the right set of people have access to it.
6. **Data Engineer :** Data engineers understand the software tools and techniques required to analyse the data to meet the desired outcomes. They know to extract, transform, load, and analyse the data at scale and implement programs for the given data models.
7. **Data Scientists :** These are individuals with deep analytical talent required to understand and work with data. They build sophisticated data models that can make predictions or carry out the required analytics. They closely work with data engineers to ensure that the data models are correctly implemented. They also choose the right data analytics approaches and ensure that the overall objectives are met.

1.2 Big Data Infrastructure Challenges

Some of the major challenges in processing Big Data are as shown in Fig. 1.2.1.

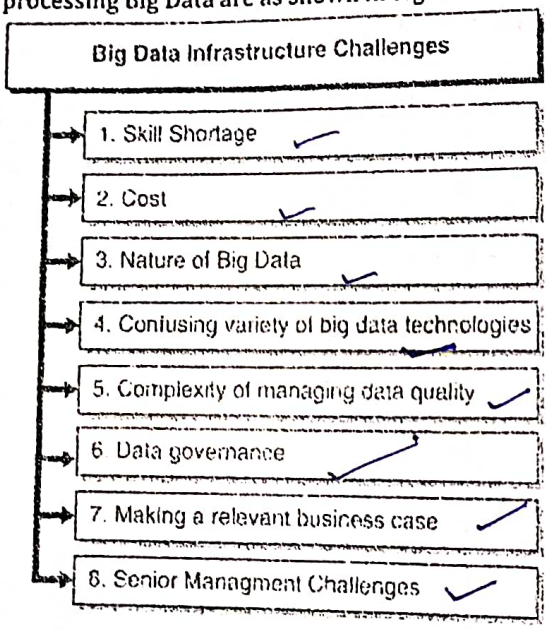


Fig. 1.2.1

1. **Skill Shortage :** The big data ecosystem is moving so fast that it is nearly impossible to keep up. New tools, capabilities, and frameworks evolve and mature in a matter of months, resulting in a skills gap that can easily impede a Big Data initiative. The demand for Big Data skills, especially in analytics, is growing significantly. Organisations are questioning how they can make long-term IT investments, leverage existing skills, and obtain new ones. Also, you require various types of skills for getting overall Big Data objectives fulfilled. Some of them are as following.
   (a) Understand the wide variety of data
   (b) Model the data correctly so as to meet the desired objectives
   (c) Build and manage software and hardware tools and techniques required for Big Data processing
   (d) Design appropriate visual interfaces and
   (e) Also communicate the findings effectively

Building a team of experts that have all the required capabilities is challenging.

2. **Cost :** Big Data ecosystem requires massive computing resources. Most big data technologies require large clusters of servers that entail long provisioning and setup cycles, resulting maintenance overhead. To make matters more complicated, the growing volume, variety, and velocity of data from either existing applications or new business requirements can result in unsustainable IT costs. Organisations need to know how to get value from big data without breaking the bank. They must be able to scale the infrastructure to manage big data while still reducing IT costs.

3. **Nature of Big Data :** Big data comes from a wide variety of sources, from legacy applications and transactional systems, to machine-generated data, mobile devices, web logs, and social media. This makes it even more difficult and inefficient to predict the required capacity. A single event can cause sudden changes in data volume and workloads. For example, a financial services organisation can experience volume fluctuations by a factor of 10 of any given day depending on market conditions.

   Organisations are challenged by Big Data's growing demands on storage capacity and computing infrastructure. Not only do organisations have to size their infrastructure, they must also determine how they will easily scale to address fluctuating storage and computing requirements. It is inefficient and cost-prohibitive for almost any organisation to size its infrastructure to support 10x volumes and let that extra capacity sit unused for 90% of the time. Additional issues include escalating infrastructure and maintenance costs due to data growth as well as ensuring adequate bandwidth to support innovation through experimentation, plus data capture and analysis.

4. **Confusing variety of big data technologies :** Data management teams have a wide range of Big Data technologies to choose from, and the various tools often overlap in terms of their capabilities. It can be easy to get lost in the variety of Big Data technologies now available on the market. Do you need Apache Spark or would the speeds of Hadoop MapReduce be enough? Is it better to store data in Cassandra or HBase? Finding the answers can be tricky. It is even easier to choose poorly, if you are exploring the ocean of technological opportunities without a clear view of what you need.

5. **Complexity of managing data quality :** Sooner or later, you will run into the problem of data integration, since the data you need to analyse comes from diverse sources in a variety of different formats. For instance, ecommerce companies need to analyse data from website logs, call-centres, competitors' website 'scans' and social media. Data formats will obviously differ and matching them can be problematic.

   Also, as you understand, Big Data isn't 100% accurate all the time. There could be unreliability issues based on the sources from where you collect Big Data and data cleaning steps that are required before you can use it. Collected data can not only contain wrong information, but also duplicate itself, as well as contain contradictions. If the data quality is poor, then it would not bring any useful insights or shiny opportunities to your precision-demanding business tasks and needs.

6. **Data Governance :** Organisation collecting a huge amount data for analytics also need to ensure proper data governance. At a high-level, data governance includes security, privacy, and ethics. There are various data collection regulations, such as GDPR, that prohibit collecting and using personal data without taking permission from the user.

   As organisations collect, store, and analyse increasing amounts of data from new and existing sources, security becomes of greater concern. They struggle to control data access, secure data assets, and protect the infrastructure. Ultimately, they are left to determine how to ensure compliance, governance, and security without compromising on agility and performance. The healthcare industry, for example, must meet HIPAA compliance, a complex undertaking that involves securing not just data but access to data via computers, printers, and copiers.

7. **Making a relevant business case :** You might think that your data analytics project, toolset, and people are cool, but the ground reality is that until you have a business case attached to your analytics project, it is hard to get senior management interested and get continuous funding. Typically, making a business case involves specifying a business goal clearly. For example, you could say that analysing the consumer purchase behaviour, you are trying to improve sales by 10%.

If you haven't built a solid business case and gathered input from powerful allies such as key business stakeholders, then chances are that you are not going to get approval for the resources you need. It is no wonder that the ability to demonstrate ROI (return on investment) is a key challenge of any data analytics project. The fact of the matter is that big data is still relatively immature and can involve a degree of experimentation, the cost of which can be very high. In order to run experiments on specific initiatives, organisations must do the undifferentiated heavy lifting that translates into a lot of time and effort. This slows down the pace of innovation and ultimately lowers the value of a big data initiative.

Organisations also tend to overemphasize the Big Data technology without understanding the context of the data and its uses for the business. There is often a ton of effort put into thinking about Big Data storage architectures, security frameworks and ingestion, but very little thought put into on boarding users and use cases. Teams need to think about who will refine the data and how. Those closest to the business problems need to collaborate with those closest to the technology to manage risk and ensure proper alignment. It is also helpful to build out a few simple end-to-end use cases to get early wins, understand the limitations and engage users and other stakeholders.

8. **Senior Management Challenges :** Some of challenges faced by senior management in an organisation with respect to Big Data are as following.
   (a) Establishing the vision for a data-driven enterprise and leading digital transformation
   (b) Driving culture change and improving data literacy
   (c) Treating data and analytics as assets and maximising their economic value
   (d) Defining a modern operating model that balances service delivery, collaboration and competency building
   (e) Applying trusted (governance, ethics, privacy, security) data and analytics in decision making
   (f) Creating a business case to support data and analytics monetisation and innovation

1.3 Big Data Processing Architectures

Big Data processing architectures are evolving from traditionally being on-premises datacentres to now on cloud computing environment. Let's learn about some of the basic architectures.

1.3.1 Data Warehouse

Definition : *A data warehouse is a central repository of information that can be analysed to make more informed decisions.*

- A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources, such as point-of-sale transactions, marketing automation, customer relationship management, and more. Data flows into a data warehouse from transactional systems, relational databases, and other sources, periodically.

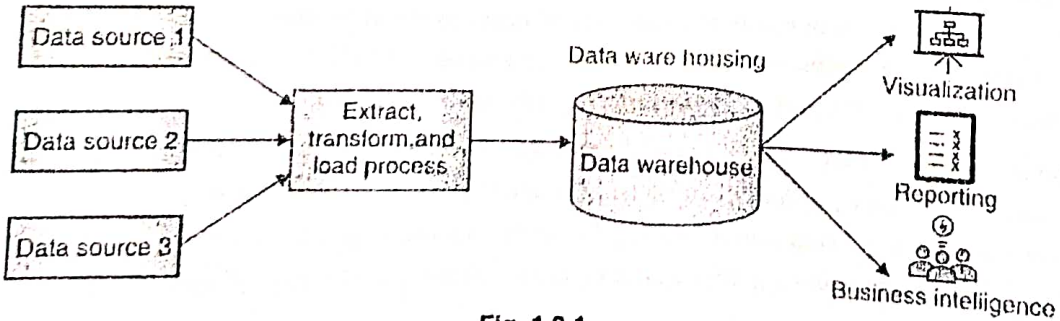


Fig. 1.3.1

- A data warehouse is suited for ad hoc analysis as well custom reporting. A data warehouse can store both current and historical data in one place and is designed to give a long-range view of data over time, making it a primary component of business intelligence. Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications.
- Data and analytics have become indispensable to businesses to stay competitive. Business users rely on reports, dashboards, and analytics tools to extract insights from their data, monitor business performance, and support decision making. Data warehouses power these reports, dashboards, and analytics tools by storing data efficiently to minimise the input and output (I/O) of data and deliver query results quickly to hundreds and thousands of users concurrently.

1.3.2 Data Warehouse Architecture

Generally speaking, data warehouses have a three-tier architecture as shown in the Fig. 1.3.2.

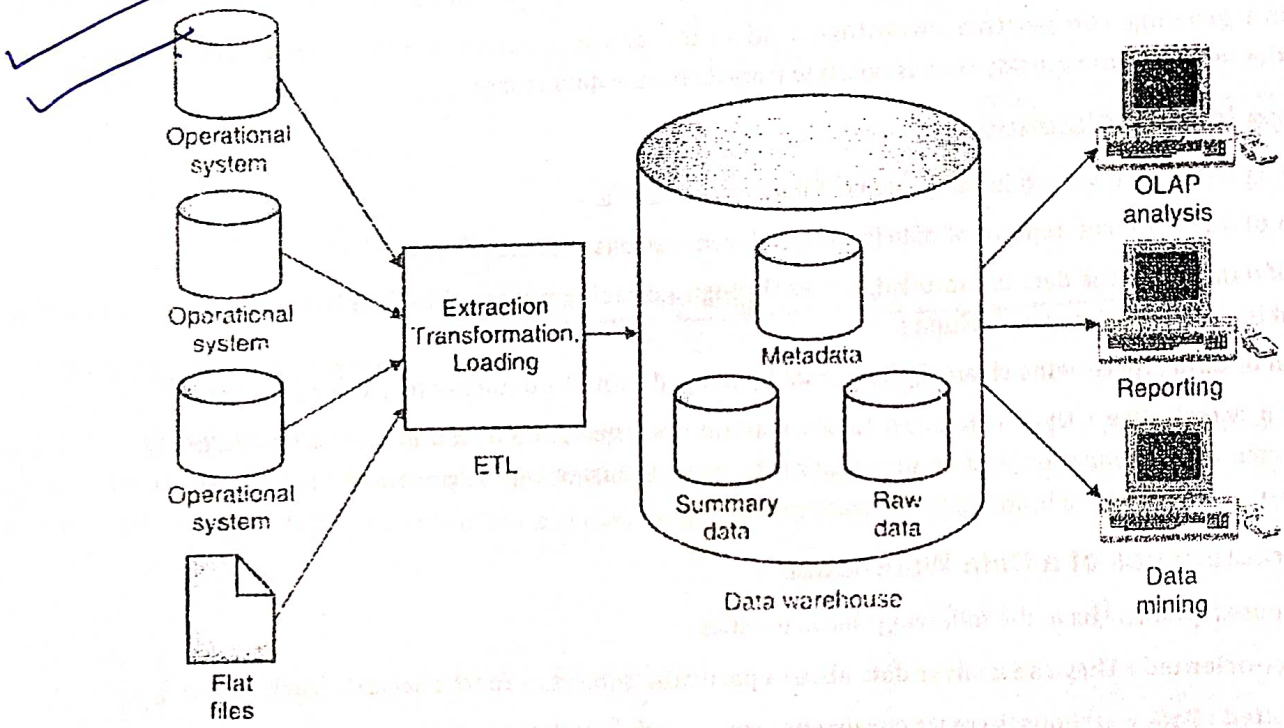


Fig. 1.3.2

1. **Bottom tier :** The bottom tier consists of a data warehouse server, usually a relational database system, which collects, cleanses, and transforms data from multiple data sources through a process known as Extract, Transform, and Load (ETL) or a process known as Extract, Load, and Transform (ELT). It is the database server where data is loaded and stored.
2. **Middle tier :** The middle tier consists of the analytics engine that is used to access and analyse the data. It consists of an OLAP (i.e. online analytical processing) server which enables fast query speeds. Three types of OLAP models can be used in this tier, which are known as ROLAP, MOLAP and HOLAP. The type of OLAP model used is dependent on the type of database system that exists.
3. **Top tier :** The top tier is represented by some kind of front-end user interface or reporting tool, which enables end users to conduct ad-hoc data analysis on their business data. It is the front-end client that presents results through reporting, analysis, and data mining tools.

1.3.3 Benefits of a Data Warehouse

A data warehouse provides a foundation for the following.

1. **Better data quality :** A data warehouse centralises data from a variety of data sources, such as transactional systems, operational databases, and flat files. It then cleanses it, eliminates duplicates, and standardises it to create a single source of the truth.
2. **Faster, business insights :** Data from disparate sources limit the ability of decision makers to set business strategies with confidence. Data warehouses enable data integration, allowing business users to leverage all of a company's data into each business decision.
3. **Smarter decision-making :** A data warehouse supports large-scale Business Intelligence functions such as data mining (finding unseen patterns and relationships in data), artificial intelligence, and machine learning—tools data professionals and business leaders can use to get hard evidence for making smarter decisions in virtually every area of the organisation, from business processes to financial management and inventory management.
4. **Gaining and growing competitive advantage :** All of the above combine to help an organisation finding more opportunities in data, more quickly than is possible from disparate data stores.

1.3.4 Steps in Data Warehousing

The following steps are involved in the process of data warehousing.

1. **Extraction of data :** A large amount of data is gathered from various sources.
2. **Cleaning of data :** Once the data is compiled, it goes through a cleaning process. The data is scanned for errors, and any error found is either corrected or excluded.
3. **Conversion of data :** After being cleaned, the format is changed from the database to a warehouse format.
4. **Storing in a warehouse :** Once converted to the warehouse format, the data stored in a warehouse goes through processes such as consolidation and summarisation to make it easier and more coordinated to use. As sources get updated over time, more data is added to the warehouse.

1.3.5 Characteristics of a Data Warehouse

- Data warehouses typically have the following characteristics.
  1. **Subject-oriented :** They can analyse data about a particular subject or functional area (such as sales).
  2. **Integrated :** Data warehouses create consistency among different data types from disparate sources.
  3. **Non-volatile :** Once data is in a data warehouse, it is stable and does not change.
  4. **Time-variant :** Data warehouse analysis looks at change over time.
- A well-designed data warehouse will perform queries very quickly, deliver high data throughput, and provide enough flexibility for end users to "slice and dice" or reduce the volume of data for closer examination to meet a variety of demands whether at a high level or at a very fine, detailed level. The data warehouse serves as the functional foundation for middleware BI environments that provide end users with reports, dashboards, and other interfaces.

1.3.6 Schemas in Data Warehouses

Schemas are ways in which data is organised within a database or data warehouse. There are two main types of schema structures, the star schema and the snowflake schema, which will impact the design of your data model.

1.3.6(A) Star Schema

This schema consists of one fact table which can be joined to a number of denormalized dimension tables. It is considered the simplest and most common type of schema, and its users benefit from its faster speeds while querying.

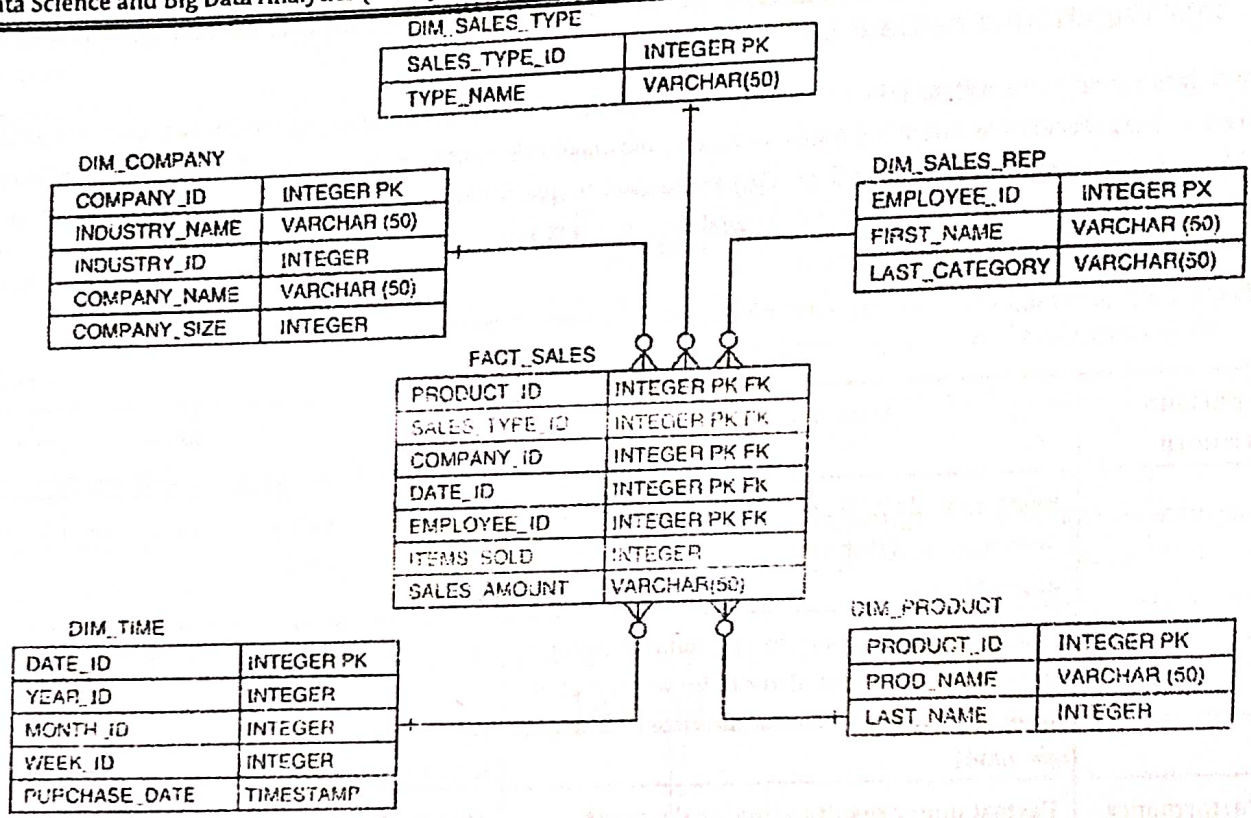


Fig. 1.3.3

1.3.6(B) Snowflake Schema

While not as widely adopted, the snowflake schema is another organisation structure in data warehouses. In this case, the fact table is connected to a number of normalised dimension tables, and these dimension tables have child tables. Users of a snowflake schema benefit from its low levels of data redundancy, but it comes at a cost to query performance.

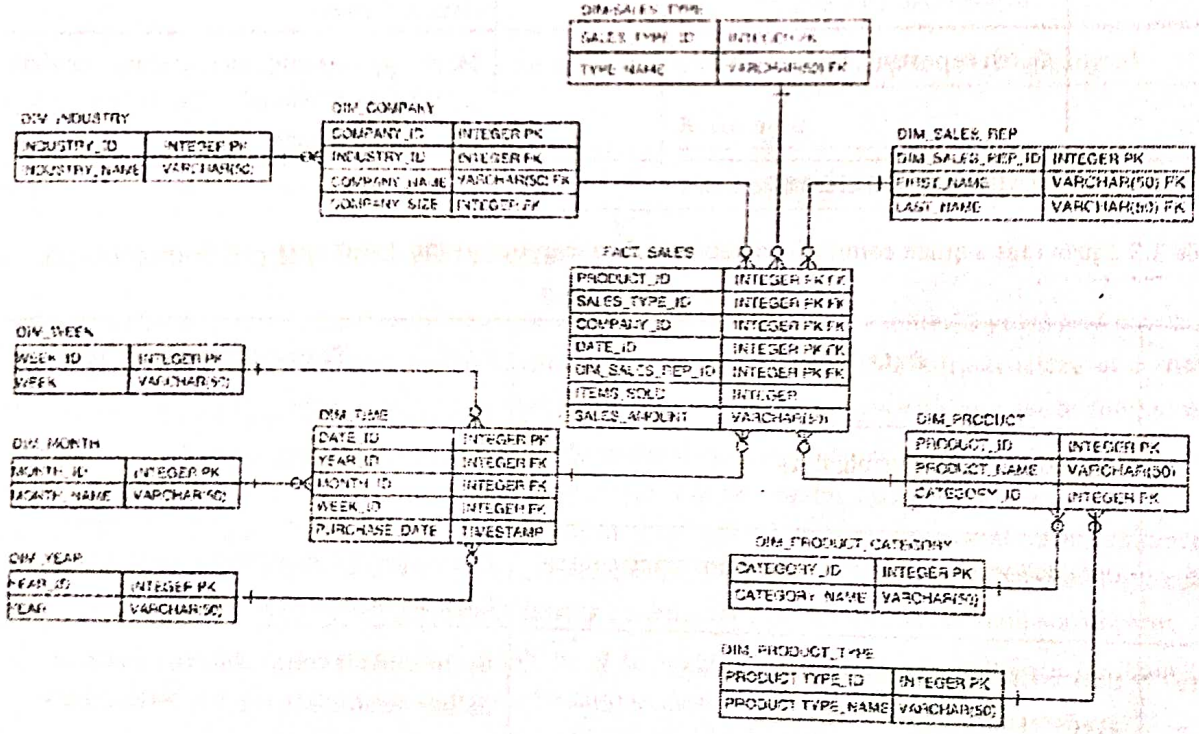


Fig. 1.3.4

1.3.7 Data Warehouse vs Data Lake

Unlike a data warehouse, a data lake is a centralised repository for all data, including structured, semi-structured, and unstructured. A data warehouse requires that the data be organised in a tabular format, which is where the schema comes into play. The tabular format is needed so that SQL can be used to query the data. But not all applications require data to be in tabular format. Some applications, like big data analytics, full text search, and machine learning, can access data even if it is 'semi-structured' or completely unstructured.

The Table 1.3.1 provides a quick comparison between data warehouse and data lake.

Table 1.3.1

Comparison Attribute	Data Warehouse	Data Lake
Data	Relational data from transactional systems, operational databases, and line of business applications	All data, including structured, semi-structured, and unstructured
Schema	Often designed prior to the data warehouse implementation but also can be written at the time of analysis (schema-on-write or schema-on-read)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using local storage	Query results getting faster using low-cost storage and decoupling of compute and storage
Data quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e. raw data)
Users	Business analysts, data scientists, and data developers	Business analysts (using curated data), data scientists, data developers, data engineers, and data architects
Analytics	Batch reporting, BI, and visualisations	Machine learning, exploratory analytics, data discovery, streaming, operational analytics, big data, and profiling

1.3.8 Data Warehouse vs Database

The Table 1.3.2 provides a quick comparison between data warehouse and database.

Table 1.3.2

Comparison Attribute	Data Warehouse	Transactional Database
Suitable workloads	Analytics, reporting, big data	Transaction processing
Data source	Data collected and normalised from many sources	Data captured as-is from a single source, such as a transactional system
Data capture	Bulk write operations typically on a predetermined batch schedule	Optimised for continuous write operations as new data is available to maximise transaction throughput

Comparison Attribute	Data Warehouse	Transactional Database
Data normalisation	Denormalised schemas, such as the Star schema or Snowflake schema	Highly normalised, static schemas
Data storage	Optimised for simplicity of access and high-speed query performance using columnar storage	Optimised for high throughout write operations to a single row-oriented physical block
Data access	Optimised to minimise I/O and maximise data throughput	High volumes of small read operations

1.3.9 Data Warehouse vs Data Mart

- A data mart is a data warehouse that serves the needs of a specific team or business unit, like finance, marketing, or sales. It is smaller, more focused, and may contain summaries of data that best serve its community of users. A data mart might be a portion of a data warehouse, too.
- The Table 1.3.3 provides a quick comparison between data warehouse and data mart.

Table 1.3.3

Comparison Attribute	Data Warehouse	Data Mart
Scope	Centralised, multiple subject areas integrated together	Decentralised, specific subject area
Users	Organisation-wide	A single community or department
Data source	Many sources	A single or a few sources, or a portion of data already collected in a data warehouse
Size	Large, can be 100's of gigabytes to petabytes	Small, generally up to 10's of gigabytes
Design	Top-down	Bottom-up
Data detail	Complete, detailed data	May hold summarised data

1.3.10 Re-Engineering the Data Warehouse

The data warehouse has been built on technologies that have been around for over 30 years and infrastructure that is at least three generations old, compared to the advancements in the current-state infrastructure and platform options. Some of the most popular and proven options to reengineer or modernise the data warehouse are as following.

1. **Re-platforming :** A very popular option is to re-platform the data warehouse to a new platform including all hardware and infrastructure. There are several new technology options in this realm, and depending on the requirement of the organisation, any of these technologies can be deployed. The choices include data warehouse appliances, commodity platforms, tiered storage, cloud computing, and in-memory technologies. Out of these options, cloud computing is the most popular one. Cloud computing offers the following benefits for re-platforming the data warehouse.
   (a) **Low total cost of ownership :** You don't need to upfront buy hardware, software, or have in-house data warehouse experts for regular operations and maintenance.

(b) **Improved speed and performance :** Cloud computing resources are highly optimised to give you performance.

(c) **Seamless self-service capabilities for business users :** Cloud data warehouse resources are easy and quick to deploy in a self-service fashion. You don't need long lead times for purchasing the hardware and commissioning them in your datacentre.

(d) **More secure data :** Cloud computing offers granular role based access control and various industry accepted security controls and practices.

(e) **Increased data storage :** Cloud computing offers nearly unlimited capacity to store data at fraction of cost.

(f) **Improved access and integration :** It is comparatively easier to get various data sources integrated with cloud storage for keeping the data warehouse up to date.

(g) **Better disaster recovery :** Cloud computing offers resiliency and disaster recovery. Several copies of your data is replicated and stored at multiple locations to protect against any disaster.

(h) **Leveraged cloud flexibility and agility :** Cloud computing offers flexibility and agility. You could up scale and down scale computing resources based on your requirements and save money by not committing to a fixed resource consumption.

2. **Platform engineering :** With advances in technology, there are several choices to enable platform engineering. This is fundamentally different from re-platforming, where you can move the entire data warehouse. With a platform engineering approach, you can modify pieces and parts of the infrastructure and get great gains in scalability and performance. The concept of platform engineering was prominent in the automotive industry where the focus was on improving quality, reducing costs, and delivering services and products to end users in a highly cost-efficient manner. By following these principles, the Japanese and Korean automakers have crafted a strategy to offer products at very competitive prices while managing the overall user experience and adhering to quality that meets performance expectations.

   Borrowing on the same principles, the underlying goal of platform engineering applied to the data warehouse can translate to :

   (a) Reduce the cost of the data warehouse

   (b) Increase efficiencies of processing

   (c) Simplify the complexities in the acquisition, processing, and delivery of data

   (d) Reduce redundancies

   (e) Minimise customisation

   (f) Isolate complexity into manageable modular environments

3. **Data engineering :** Data engineering is a relatively new concept where the data structures are reengineered to create better performance. In this exercise, the data model developed as a part of the initial data warehouse is often scrubbed and new additions are made to the data model. Typical changes include the following.

   (a) **Partitioning :** A table can be vertically partitioned depending on the usage of columns, thus reducing the span of I/O operations. This is a significant step that can be performed with minimal effect on the existing data and needs a significant effort in ETL and reporting layers to refresh the changes. Another partition technique already used is horizontal partitioning where the table is partitioned by date or numeric ranges into smaller slices.

   (b) **Colocation :** A table and all its associated tables can be colocated in the same storage region. This is a simple exercise but provides powerful performance benefits.

   (c) **Distribution :** A large table can be broken into a distributed set of smaller tables and used. The downside is when a user asks for all the data from the table, you have to join all the underlying tables.

(d) New data types : Several new data types like geospatial and temporal data can be used in the data architecture and current workarounds for such data can be retired. This will provide a significant performance boost.

(e) New database functions : Several new databases provide native functions like scalar tables and indexed views and can be utilised to create performance boosts.

Though there are several possibilities, data engineering can be done only if all other possibilities have been exhausted. The reason for this is there is significant work that needs to be done in the ETL and reporting layers if the data layer has changes. This requires more time and increases risk and cost. Therefore, data engineering is not often a preferred technique when considering reengineering or modernizing the data warehouse.

1.4 Shared-Everything and Shared-Nothing Architecture

In the world of massively parallel processing (MPP) architectures for analytics databases, there have been two predominant approaches – "shared-nothing", and "shared-everything". Let's learn about them.

1.4.1 Shared-Nothing Architecture

- In a shared-nothing environment, each server operates independently, and controls its own memory and disk resources. Data is partitioned among the servers, and the workload is distributed such that each machine operates on its own data, without sharing hardware resources with other machines in the grid. The Fig. 1.4.1 shows a shared-nothing architecture.

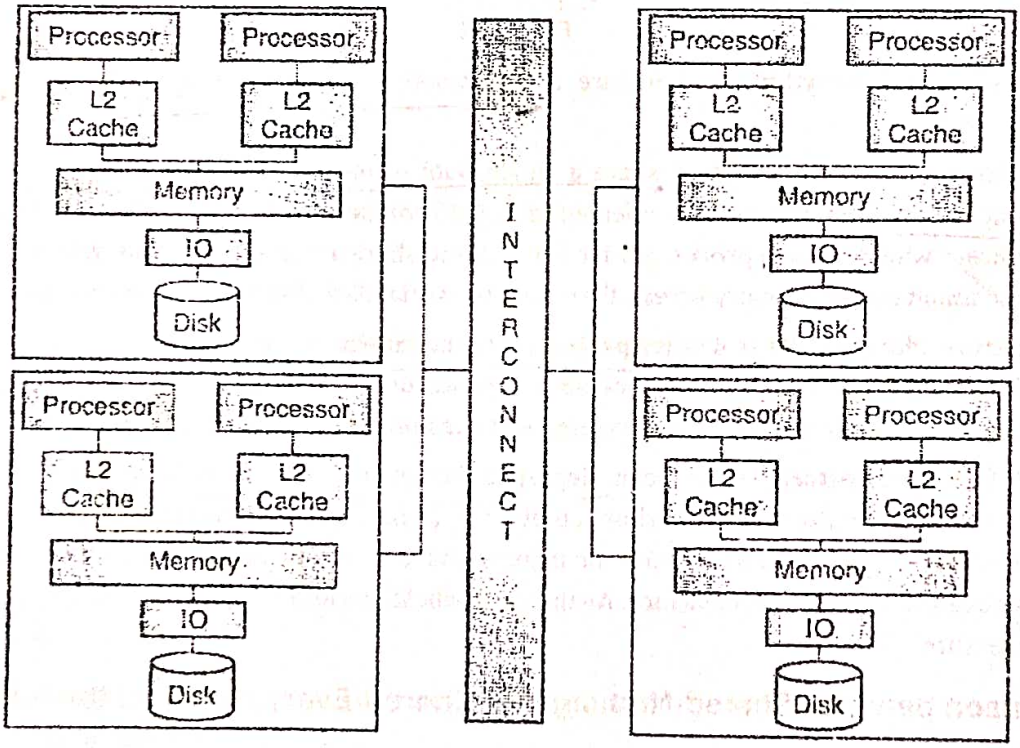


Fig. 1.4.1

- The key feature of shared-nothing architecture is that the operating system not the application server owns responsibility for controlling and sharing hardware resources. In a shared-nothing architecture, a system can assign dedicated applications or partition its data among the different nodes to handle a particular task. Shared-nothing architectures enable the creation of a self-contained architecture where the infrastructure and the data coexist in dedicated layers.

1.4.2 Shared-Everything Architecture

In a shared-everything environment, all servers access the same shared store, and each workload has access to th store, as well as the computing resources of all servers in the grid. The Fig. 1.4.2 shows a shared-everythin architecture.

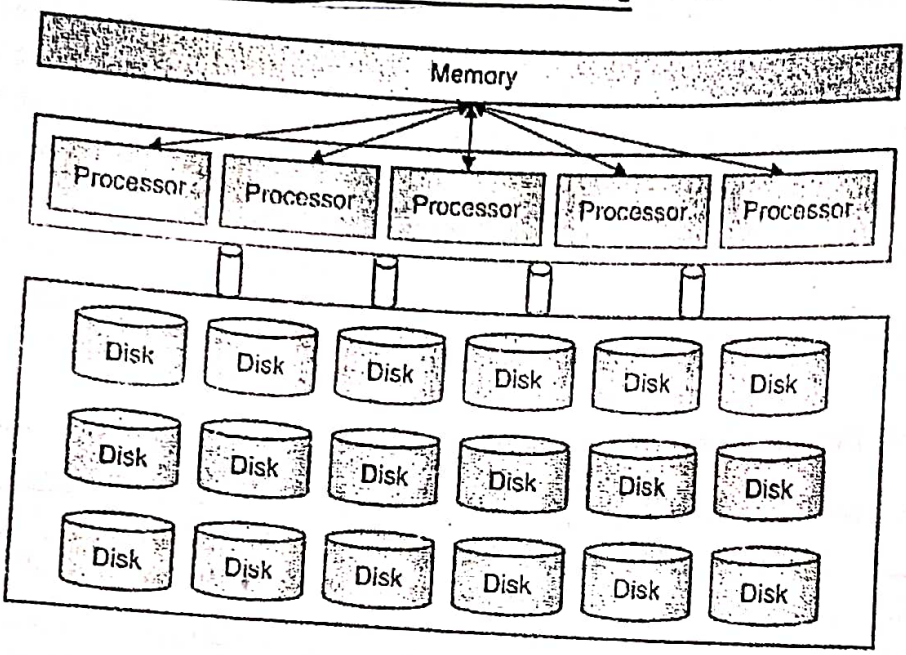


Fig. 1.4.2

- Two variations of shared-everything architecture are symmetric multiprocessing (SMP) and distributed shared memory (DSM).
- In the SMP architecture, all the processors share a single pool of memory for read-write access concurrently and uniformly without latency. Sometimes this is referred to as uniform memory access (UMA) architecture. The drawback of SMP architecture is when multiple processors are present and share a single system bus, which results in choking of the bandwidth for simultaneous memory access, therefore, the scalability of such system is very limited.
- The DSM architecture addresses the scalability problem by providing multiple pools of memory for processors to use. In the DSM architecture, the latency to access memory depends on the relative distances of the processors and their dedicated memory pools. This architecture is also referred to as non uniform memory access (NUMA) architecture.
- Both SMP and DSM architectures have been deployed for many transaction processing systems, where the transactional data is small in size and has a short burst cycle of resource requirements. Data warehouses have been deployed on the shared-everything architecture for many years, and due to the intrinsic architecture limitations, the direct impact has been on cost and performance. Analytical applications and Big Data cannot be processed on a shared-everything architecture.

1.4.3 Comparison between Shared-Nothing and Shared-Everything Architecture

The Table 1.4.1 provides a quick comparison between shared-nothing and shared-everything architecture.

Table 1.4.1

Comparison Attribute	Shared-Nothing	Shared-Everything
Focus	Maximise Performance	Maximise Resource Utilisation
Data Duplicated?	Yes	No
Data Partitioning required?	Yes	No

Comparison Attribute	Shared-Nothing	Shared-Everything
Cost	High	Low
Maintenance	Low	High
Scaling	Easy	Difficult

1.5 Data Analytics Life Cycle (Big Data Analytics)

Q. Explain different phases of data analytics life cycle. **SPPU - Aug. 18, 6 Marks**

Q. Explain Data Analytic Life cycle. **SPPU - Dec. 18, 8 Marks**

Q. Draw Data Analytics Lifecycle & give brief description about all phases. **SPPU - May 19, 5 Marks**

Q. Demonstrate the overview of Data Analytics Life Cycle. **SPPU - Oct. 19, 5 Marks**

The data analytics life cycle broadly has six phases. Each of these phases are worked through iteratively with the previous phase before moving to the next phase.

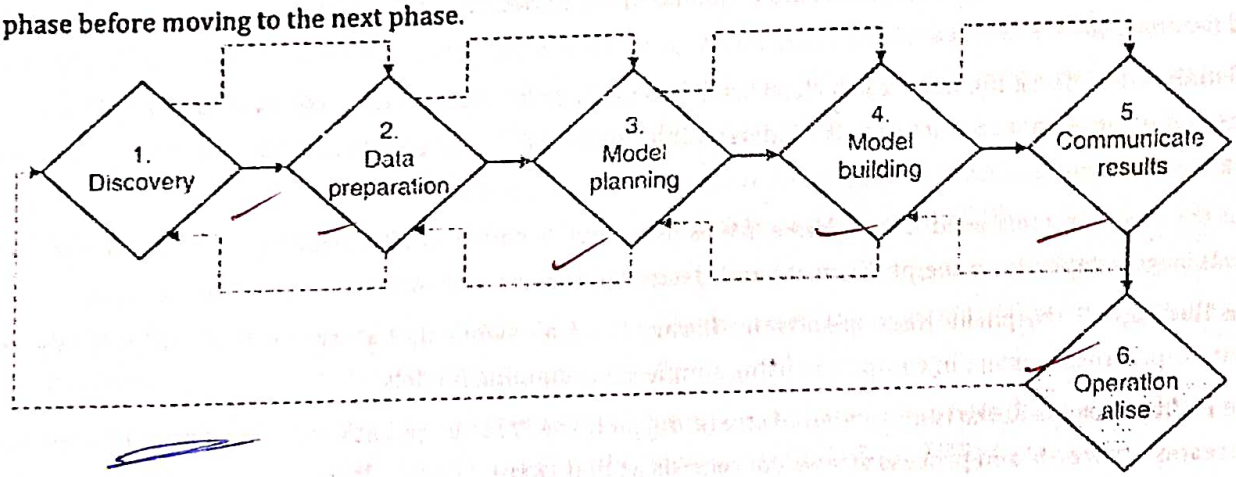


Fig. 1.5.1 : Data Analytics Life Cycle

1.5.1 Phase 1 : Discovery

In the Discovery phase, the data science team

1. Learns about the business problem to solve,
2. Investigates the problem,
3. Develops context and understanding,
4. Examines the available data sources and
5. Formulates the initial hypothesis

The team learns about the business domain in which the problem is to be solved. It assesses the resources available for the project and carries out the feasibility analysis. It spends time in framing the right problem.

Definition : *Framing is the process of stating the analytics problem to be solved.*

As part of the framing activity, the main objectives of the project are ascertained and the success criteria for the project is clearly defined. It also develops the initial hypothesis that can later be substantiated with the data.

1.5.2 Phase 2 : Data Preparation

The data preparation phase explores, pre-processes and conditions the data before modelling and analysis could be carried out. In this phase, the following activities are carried out.

1. **Preparing the analytics environment :** In this step, an isolated workspace is created in which the team can explore the data without interfering with the live data. The data from various data sources is collected in the isolated workspace.
2. **Perform ETL process :** ETL stands for Extract, Transform and Load. In this step, the raw data is extracted from the datastore, transformed as deemed right (removing noise, outliers, and biases from data) and then loaded into the datastore again for analysis.
3. **Learn about the data :** Once the ETL process is complete, the team spends time in learning about the data and its attributes. Understanding the data itself is the key to building a good data model in the subsequent phase.
4. **Data conditioning:** In this step, the data is further cleaned and normalized by performing further transformations as required. The data from several sources could be joined or combined as required. The actual data attributes that would be used for analytics are decided.
5. **Data visualisation:** Once the data is in a clean state and ready to be analysed, it is a good idea to visualise it to identify patterns and explore data characteristics. Understanding patterns about the data enables building a perspective about the data model.

Some of the common tools used in this phase are as following. Note here that the following list is not exhaustive. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **Apache Hadoop :** The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
2. **Apache Kafka :** Apache Kafka is a distributed streaming platform. You can publish and subscribe to streams of records, store streams of records and process streams of records as they occur.
3. **Alpine Miner :** Alpine Miner provides a graphical interface for creating analytics workflows and is optimised for fast experimentation, collaboration, and an ability to work within the database itself.
4. **OpenRefine :** OpenRefine is a powerful tool for working with messy data. It cleans the data and transforms it from one format into another.

1.5.3 Phase 3 : Model Planning

In this phase, the team explores and evaluates the possible data models that could be applied to the given datasets to get the desired results. The team can try several models before finalising. Some of the major activities carried out in this phase are as following.

1. **Data Exploration :** The team spends time in understanding the available data and the various patterns and relationships amongst its attributes. The team could consult subject matter experts, stakeholders, analysts, and others who might have a viewpoint on how the data should be interpreted and examined.
2. **Model Selection :** The goal of this activity to choose an analytical technique based on the given dataset and the desired outcome. Based on the type of data (structured, semi-structured or unstructured) different techniques could be chosen and applied.

Some of the common tools used in this phase are as following. Note here that the following list is not exhaustive. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **R** : R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
2. **SQL Server Analysis Services** : SQL Server Analysis Services supports tabular models at all compatibility levels, multidimensional models, data mining, and Power Pivot for SharePoint. It provides an analytical data engine used in decision support and business intelligence (BI) solutions, providing the analytical data for business reports and client applications such as Excel, Reporting Services reports, and other third-party BI tools.
3. **SAS/ACCESS** : It provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC and OLE DB. You can access the most popular databases on common platforms without detailed knowledge of the database or SQL.

1.5.4 Phase 4 : Model Building

In this phase, the team starts to build the data analytics model. The available dataset is divided into

1. Training dataset, 2. Testing dataset and 3. Production dataset

The training dataset is used to train (design) the model. Once the team is confident about the model, it tests the model using the testing dataset. Once the testing is complete, the model is ready to be used in the production (go live). The production dataset or new datasets could be applied to it to get the desired results.

Some of the common tools used in this phase are as following. Note here that the following list is not exhaustive. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **R :** R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
2. **GNU Octave :** It is a scientific programming language with powerful mathematics-oriented syntax with built-in plotting and visualization tools. It is free software that runs on GNU/Linux, macOS, BSD, and Windows.
3. **WEKA :** It is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
4. **Python :** It is a is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
5. **Commercial software :** Apart from the free and open source tools, there are various commercial software available for data analytics such as SAS Enterprise Miner, SPSS Modeler, MATLAB, and Alpine Miner. You could choose the one that suits your requirements and matches your project budget.

1.5.5 Phase 5 : Communicate Results

Q. Why communication is important in data analytics lifecycle projects ? **SPPU - May 19, 8 Marks**

After building, testing, and executing the model, the team compares the outcome with the pre-established success criteria. The results are validated and could be statistically proven. The team then articulates the findings and documents the results. The findings are communicated to the project stakeholders.

Note here that the model building exercise could be unsuccessful. The findings are still documented and reported before the team goes on to try and build another model. Recall from the data analytics life cycle diagram that each phase is an iterative process and works with the previous phase.

1.5.6 Phase 6 : Operationalise

In the final phase, the model is deployed in the staging environment before it goes live on a wider scale. The staging environment is very similar to the production environment. The idea is to ensure that the model sustains the performance requirements and other execution constraints and any issues are identified before the model is deployed in the production environment. If any changes are required, they are carried out and tested again.

The project outcome is shared with the key stakeholders such as

1. **Business user :** The business user ascertains the benefits and implications of the project findings.
2. **Project sponsor :** The project sponsor asks questions around ROI (return on investment) and any potential risks to maintaining the project.
3. **Project manager :** The project manager determines if the project was timely completed and the goals were met.
4. **Business intelligence analyst :** The business intelligence analyst determines if any of the reports or dashboards needs to be changed to accommodate the new findings.
5. **Database Administrator :** The database administrator needs to plan for backup of datasets and any other code that was written to be run on the database for the analytics project.
6. **Data Engineer :** The data engineer needs to share the code; version control it and maintain it. Any issues or bugs found in the code should be fixed.
7. **Data Scientist :** The data scientist could explain the model to her peers and other stakeholders. She also documents the model and how it was implemented.

1.6 Data Science – The Big Picture

Definition : *Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.*

Data science as a concept unifies several domains such as statistics, data analysis, informatics, and their related methods in order to understand and analyse real-world phenomena with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights on business data. Data science is heavy on computer science and mathematics. Data science is used in business functions such as strategy formation, decision making and operational processes. It touches on practices such as artificial intelligence, analytics, predictive analytics and algorithm design.

1.6.1 Business Intelligence (BI) vs Data Science

It is easy to get confused between Business Intelligence (BI) and Data Science or Big Data analytics.

✍ ***Definition :*** *BI enables an organisation to gain insight into its performance by analysing its business processes and information systems.*

- Note here that the organisation mostly uses internally available data for business intelligence purpose. This could be sales report, user feedback, customer surveys, website traffic analysis or typical business applications such as ERP and CRM. BI provides periodic business reports to
  1. Help gain business insights
  2. Drive business growth and performance and
  3. Fix business issues
- Whereas, data analytics tends to collect and use data over a large period of time to build future predictions. The data is collected from various sources (both externally and internally) and prediction models are built to forecast the things of interest.
- The Table 1.6.1 summarises the comparison between business intelligence and data analytics.

Table 1.6.1

Comparison Attribute	Business Intelligence	Data Analytics
Nature of analysis	Illustrative	Predictive
Time horizon	Short	Long
Data Formats used	Mostly structured data	Mostly unstructured data
Tools used	Database and queries	Analytics tools and techniques
Nature of questions answered	Close ended (when, what, where)	Open ended (how, why, what if)
Complexity	Low	High
Skills required	Low	High
Source of Data	Mostly Internal	Mostly External

1.6.2 Relationship between Data Science and Information Science

✍ ***Definition :*** *Information science is a discipline that deals with the processes of storing and transferring information.*

- It brings together concepts and methods from disciplines such as library science, computer science and engineering, linguistics, and psychology in order to develop techniques and devices to aid in information handling. Information handling often requires collection, organisation, storage, retrieval, interpretation, and ultimately the use of information.
- Note here that the key difference between information science and data science is that information science is about information handling and not running analytics on top of the available information (or data). Though you can say that for a meaningful or successful data analytics, you need to anyways handle the data correctly but distinctively those are two independent fields. Data science and information science are distinct but complimentary disciplines.

- Data science is heavy on computer science and mathematics. Information science is more concerned with areas such as library science, cognitive science and communications. Data science is used in business functions such as strategy formation, decision making and operational processes. It touches on practices such as artificial intelligence, analytics, predictive analytics and algorithm design. Information science is used in areas such as knowledge management, data management and interaction design.
- The Table 1.6.2 summarises the comparison between information science and data science.

Table 1.6.2

Comparison Attribute	Information Science	Data Science
Analytics carried out?	No	Yes
Use of Mathematics	Low	High
Computing resource requirements	Small	Large
Dependency on data	Low	High
Human judgment involved	To a certain extent	Minimal

1.7 Introduction to Machine Learning

- What do you mean by learning ? What do you mean when you tell someone to "learn" something? The dictionary meaning of the word learn is "to gain knowledge or understanding of or skill in by study, instruction, or experience".
- You, as a reader or student of a particular course, learn something and then become equipped or capable of carrying out various tasks based on what you learnt. You might have heard about Bloom's Taxonomy. Bloom's Taxonomy is a classification of the different objectives and skills that learners could achieve out of a particular learning or a course. Fig. 1.7.1 outlines Bloom's Taxonomy.

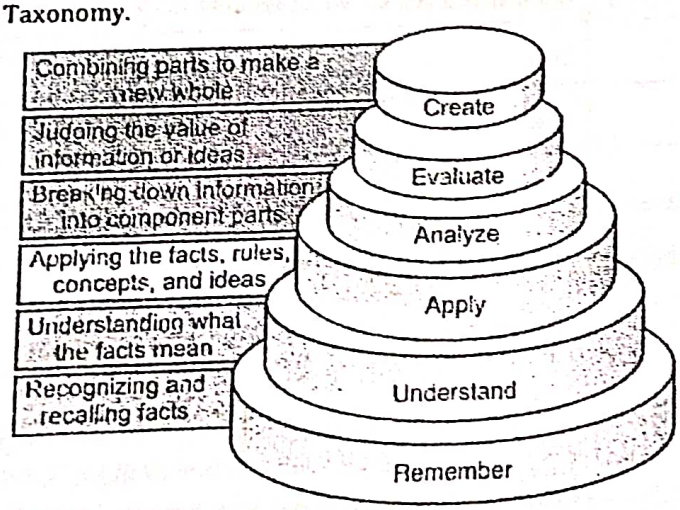


Fig. 1.7.1 : Bloom's Taxonomy

- So, as you understand, learning helps to gain skills and carry out various tasks. Like human beings, animals and other organisms also learn which helps them to carry out tasks required as per their lifecycle.

- You would also agree that, in general, the more you learn (through experience, courses, instructions, manuals, training, workshop or anything else) the more capable (and performant) you become at carrying out tasks based on learning objectives.

  So, if

  E = Experience

  P = Performance for a given task

  T = Task

- Then, you can mathematically say that performance at a given task is directly proportional to the experience. Isn't it?

  $P\{T\} \alpha E$

- As humans and other living organisms learn, similarly, it was thought, can machines made to learn and carry out tasks at a decent performance level?
- One of the most notable work towards Machine Learning (ML) was done by Alan Turing in 1950, yes 70 years ago! In his 1950 paper, "Computing Machinery and Intelligence," Alan Turing asked, "Can machines think?" (See http://www.csee.umbc.edu/courses/471/papers/turing.pdf for the full paper.).
- The paper describes the "Imitation Game", which involves three participants a human acting as a judge, another human, and a computer that is attempting to convince the judge that it is human. The judge would type into a terminal program to "talk" to the other two participants. Both the human and the computer would respond, and the judge would decide which response came from the computer.
- If the judge couldn't consistently tell the difference between the human and computer responses, then the computer won the game.
- The test continues today in the form of the Loebner Prize, an annual competition in artificial intelligence. The aim is simple enough : Convince the judges that they are chatting to a human instead of a computer chat bot program.
- In 1959, Arthur Samuel defined machine learning as, "[A] Field of study that gives computers the ability to learn without being explicitly programmed." Samuel is credited with creating one of the self-learning computer programs with his work at IBM. Samuel is widely known for his work in artificial intelligence.
- One the most common definition of Machine Learning came from Tom M. Mitchell, the Chair of Machine Learning at Carnegie Mellon University. It is as following.

✍ ***Definition :*** *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E.*

- Hence,

✍ ***Definition :*** *Machine Learning is the study of computer algorithms and programs that automatically improve their performance, for a given set of tasks, with increase in their experience.*

- Machine learning is a branch of artificial intelligence (making machines intelligent). Machine learning uses various statistical (mathematical) models to learn from collected data and then uses those trained models to carry out various tasks such as predictions and classifications.

1.7.1 How does Machine Learning Work ?

The Fig. 1.7.2 outlines how machine learning works at a high-level.

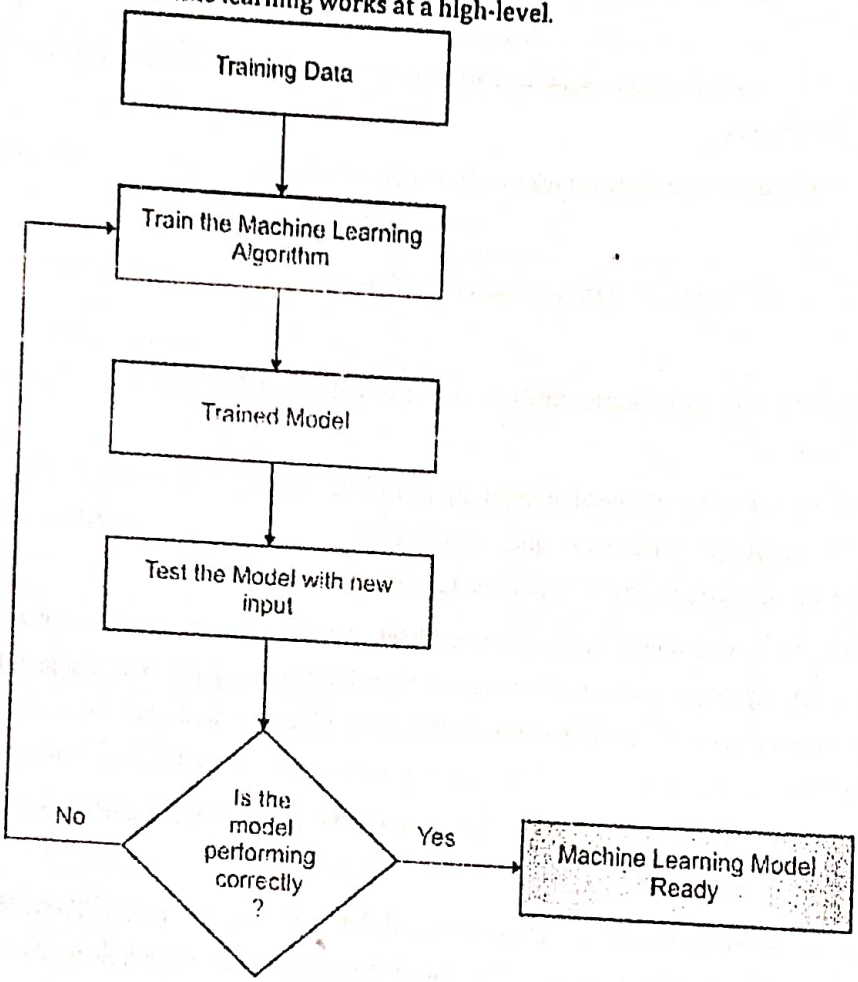
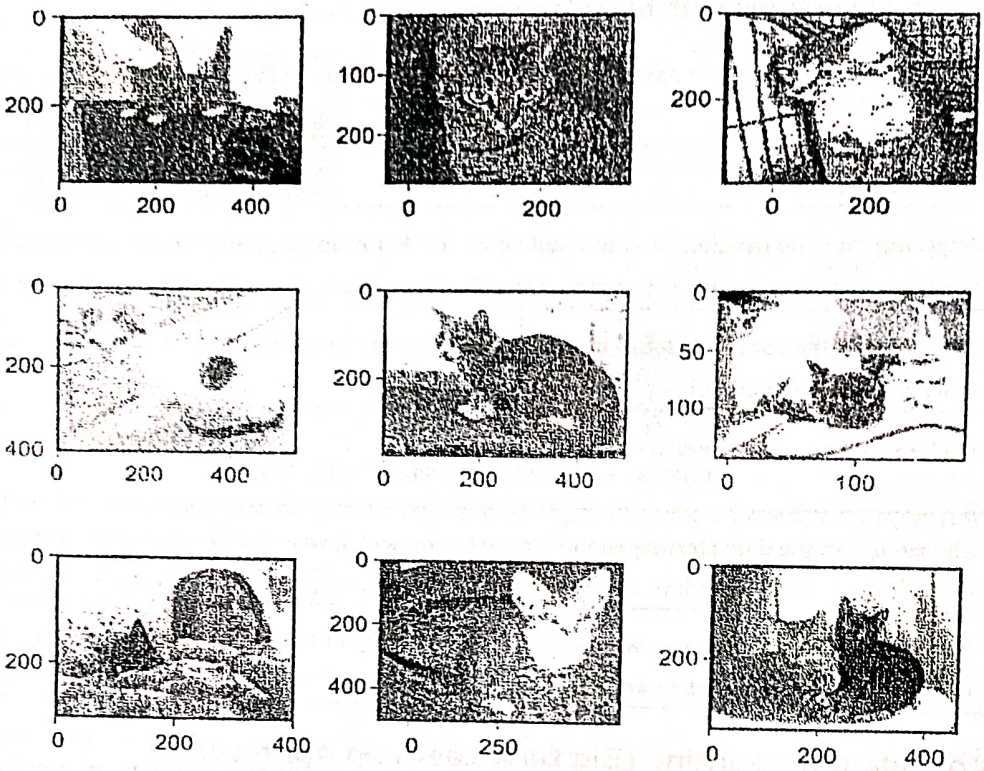


Fig. 1.7.2 : Machine Learning Model

- I am sure you would have made several predictions many a times in life already, isn't it? Whether it is going to rain, whether a team is going to win, whether the stock market would go up or down and likewise.
- Similarly. you can identify a cat pretty confidently even if it is of different shape and size and is in almost any position (sleeping, jumping, turning upside down, running, eating or whatever).
- How do you do these things? You can do this because you have some sort of mental model trained from your past experiences (and learning's).
- Based on those experiences, you can easily predict facts or identify things. If it occurs to you that what you predicted or identified was not right, you take that feedback and update your mental model so that you can do a better job next time.
- Probably, you are less likely to make the same mistake. Unfortunately, machines aren't as intelligent as you are, but we take the similar approach to make them intelligent!
- Machine learning is all about using a huge volume of training (sample) data (also called as big data analysis) and then using that data to build models that can carry out a set of tasks when a similar but new input is provided.
- For example, if you want a machine to automatically identify a cat from various animal pictures fed to it, you first train the machine to recognise different types of cats and also in various positions and environments.
- For example, assume that the following picture has the training data for identifying cats in a picture.



- Once the model has learnt "enough", it can then successfully carry out a task such as identifying a cat in a picture that was not in the training data.
- For example, it can identify the cat in the following picture.

- That's a very high-level view of how machine learning works. However, there are many more complications and effort required for building and training a complex machine learning model for real-life applications. But, a general high-level understanding of how machine learning works would go a long way to help you build the complex concepts later on.

1.7.2 Key Terms Associated with Machine Learning

Let's understand some of the key terms associated with machine learning that you would commonly encounter.

Table 1.7.1

Term	Description
Features or Attributes	Anything that you can measure and build data for. For example, the typical length of various animals. Feature could be numeric, set of characters, Boolean values, or anything else that describes the data.
Training Dataset	(Big data) or the complete set of sample data (training examples) based on which you would train your machine learning model.
Training Example	Each example (or data point) within a dataset.
Testing Dataset	The set of sample data (testing examples) based on which you could test your trained machine learning model for its correctness and determining if further training or adjustment is required to the model.
Target variable	The feature or value that you want to predict or identify or the output that the trained machine learning model should produce when an input is fed to it.

1.8 Types of Machine Learning (Big Data Learning Approaches)

As a human learn in various ways (auditory, visual, kinesthesis), so does machines. You also understand that there could be different types of data formats and various analytics and learning requirements.

At a high-level, machine learning algorithms could be categorized as following :

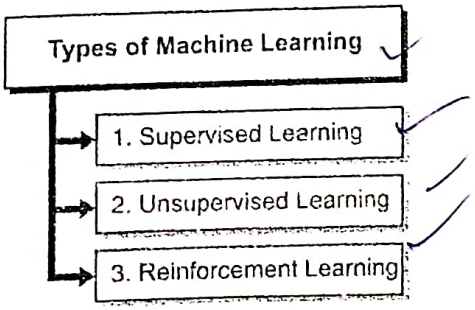


Fig. 1.8.1 : Types of Machine Learning

1.8.1 Supervised Learning

Definition : Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.

- In supervised learning, the machine learning algorithm is trained on labelled data (the training dataset has both input as well as output). This is very similar to you teaching a toddler by showing a picture of an apple and saying, "Look kid, this is an apple". Fig. 1.8.2 provides a high-level outline of how supervised machine learning algorithm work.

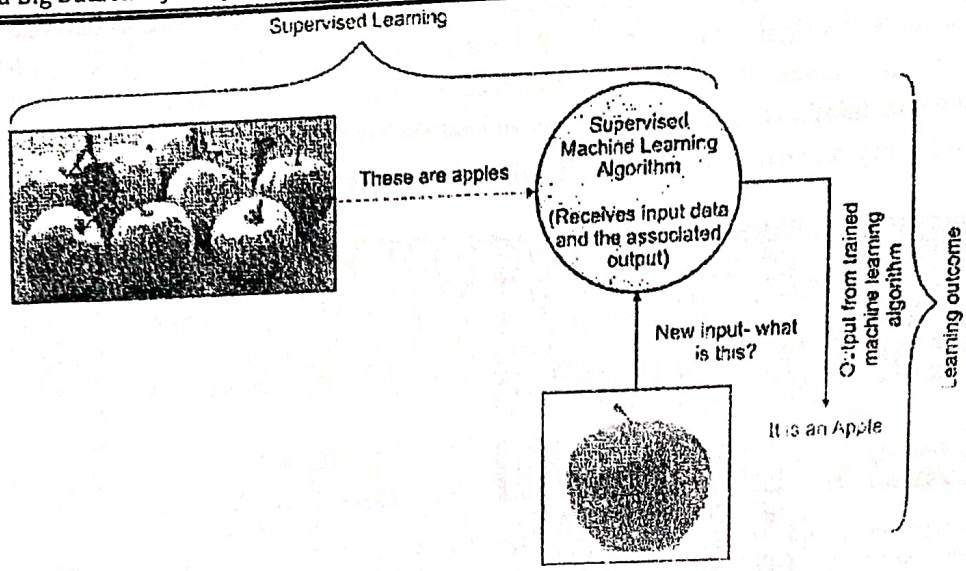


Fig. 1.8.2 : Supervised Machine Learning

- The training data set has various examples that contain both input (features) as well as output (target). Supervised learning is typically used for classification or predicting a particular value (regression). For example, classifying an animal picture as a cat or a dog or whether it is likely to rain. Some of the common supervised learning algorithms are k-Nearest Neighbours, Linear regression, Logistic regression, Naïve Bayes, Support Vector Machines, and Decision Trees.

1.8.2 Unsupervised Learning

✍ ***Definition*** *: Unsupervised learning algorithms take a set of data that contains only inputs and automatically find structure in the data in order to group or arrange them in a cluster.*

- In unsupervised learning, once the model is given a training dataset, it automatically finds patterns and relationships in the dataset by creating clusters (or groups) within it. Unsupervised learning algorithms, however, cannot label the cluster or determine what those created clusters might mean.
- For example, if you feed a training dataset having thousands of pictures of cats, dogs, and monkeys, an unsupervised learning algorithm can potentially create three different clusters – one for each animal but it cannot tell you which cluster is what. It is up to the observer or human to make the sense out of the created clusters of data.
- When you feed a new input to such an algorithm, it can place it in one of the already created clusters or could create a new group if it is dissimilar to any of the already created clusters.
- Fig. 1.8.3 provides a high-level outline of how unsupervised machine learning algorithm work.

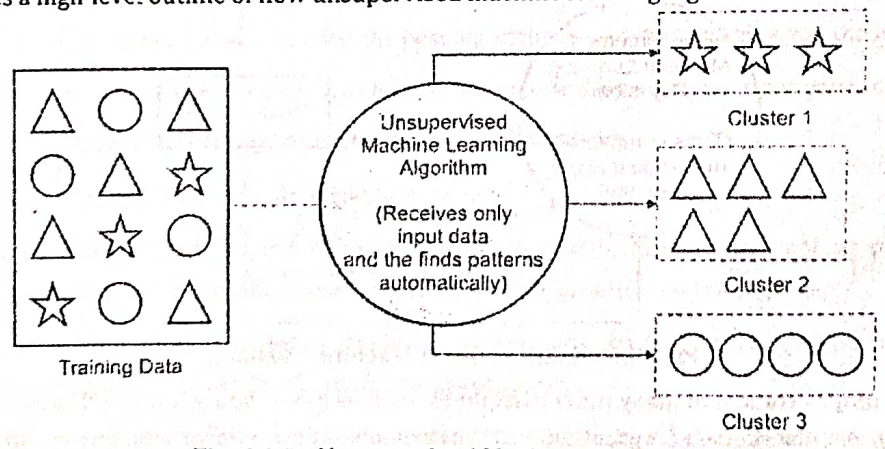


Fig. 1.8.3 : Unsupervised Machine Learning

- Thus, unsupervised learning algorithms learn from the training dataset that has not been explicitly labelled and neither has the output (or target) associated with it. Unsupervised learning algorithms are commonly used to find patterns of data and build recommendation engine. For example, you would have commonly seen various recommendations based on your purchase history on e-commerce portals or streaming portals such as YouTube based on genre of videos you have watched.

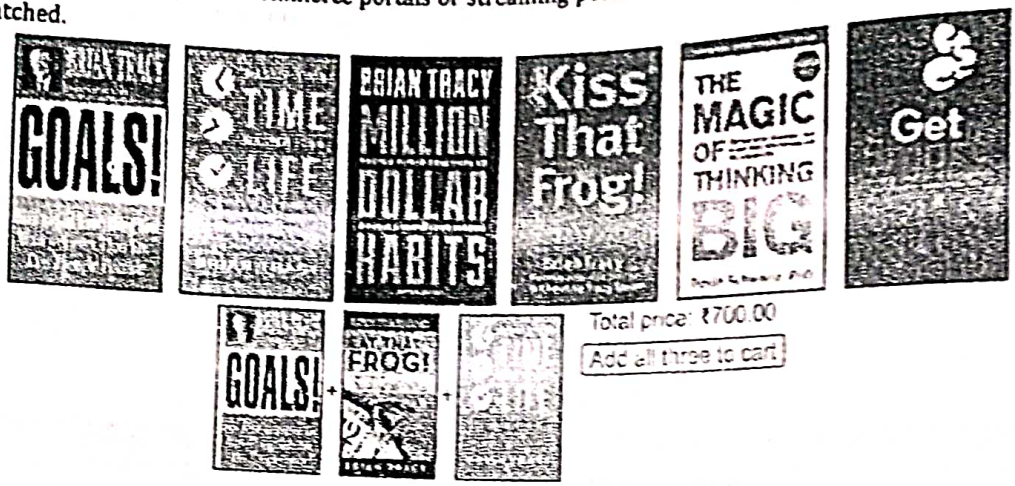


Fig. 1.8.4

- Some of the common unsupervised learning algorithms are k-Means clustering, Expectation maximization, Hidden Markov Model, DBSCAN, and Parzen window.

1.8.3 Reinforcement Learning

- Treat reinforcement learning as learning from mistakes. The reinforcement learning algorithms work very similar to how you learn by yourself without any guidance basically through hit and trial. When you get something right, you get a reward, you feel happy, and you move ahead. When you get something wrong, you get a penalty, you take a step back, and then you try to avoid the incorrect path while exploring another correct path.

Definition : *Reinforcement machine algorithms improves upon themselves and learn from new situations using a trial-and-error method.*

- The favourable outputs are encouraged, or 'reinforced', and non-favourable outputs are discouraged or 'punished'.
- Fig. 1.8.5 outlines how reinforcement learning works at a high-level.

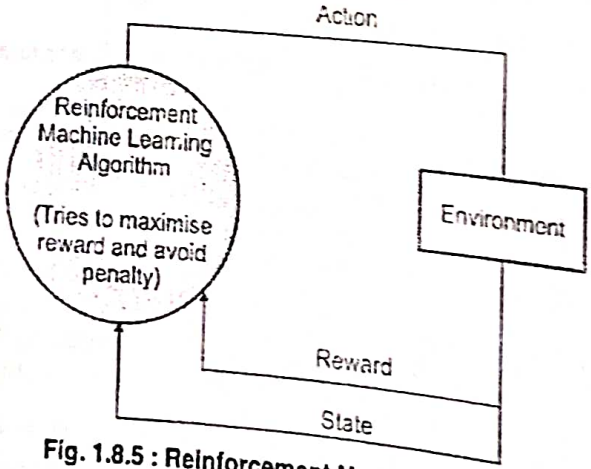


Fig. 1.8.5 : Reinforcement Machine Learning

- Reinforcement learning is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics, and genetic algorithms. It is not used frequently for machine learning problems such as classification or prediction.

- Many reinforcement learning algorithms use dynamic programming techniques. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

Comparison between Supervised and Unsupervised Learning

The Table 1.8.1 provides a quick comparison between supervised and unsupervised learning.

Table 1.8.1

Comparison Attribute	Supervised Learning	Unsupervised Learning
Training Dataset Contains	Both input and output	Only input
Used for	Classification and Prediction	Finding patterns and understanding data
Training Data	Is Labelled	Not labelled
Number of targets	Known beforehand	Not known
Feedback from user	Provided	Not provided
Complexity	High	Low

1.8.4 How to Choose the Right Machine Learning Algorithm?

- So, how do you choose which machine learning type to go with? It is simple – if you need to predict a target value, then you should choose supervised learning.
- For example, if you want to find out chances of raining tomorrow, a team winning a tournament, a picture being an apple or a cat, or problems such as these, then you should choose supervised learning method.
- However, if you are not solving a prediction or classification problem and your goal is to group the data and find "interesting" patterns, then you are better of using unsupervised learning.
- Once you have decided which type of machine learning algorithm you need, you then need to choose a suitable algorithm from that machine learning type.
- Note here that there is no single answer to what the best algorithm is or what will give you the best results.
- You will need to try different algorithms and carefully evaluate how they perform as per your requirements.

Comparison between Artificial Intelligence and Machine Learning

- Machine learning is often confused with Artificial Intelligence.
- Without going too much into details, understand that artificial intelligence is about making smart machines to solve complex problems similar to humans. Machine learning is a subset of Artificial Intelligence.
- When you interact with Google Assistant or Apple Siri or Alexa to check weather conditions or ask it to play your favourite song, you experience an Artificially built Intelligent system.
- The Table 1.8.2 provides a quick comparison between Artificial Intelligence and Machine Learning for your reference.

Table 1.8.2

Comparison Attribute	Machine Learning	Artificial Intelligence
Focus	Learn from data	Solve complex problems
Complexity	Low	High
Scope	Narrow	Broad
Human interaction	Minimal	High

Comparison between Data Mining and Machine Learning

- Data mining is a process used by companies to turn raw data into useful information.
- By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.
- However, there are significant differences between data mining and machine learning.
- The Table 1.8.3 provides a quick comparison between data mining and machine learning.

Table 1.8.3

Comparison Attribute	Machine Learning	Data Mining
Building a trained model	Required	Not required
Human effort required	Only for building model	For extracting information from data
Use of specific algorithms	Frequent	Rare
Accuracy	High	Low
Tasks carried out by	Machines	Humans
Self-learning	Yes	No

1.9 Statistical Learning

Machine learning is heavily based on statistics. You would learn about several statistical learning tools in Unit 2 under "Need of Statistics in Data Science and Big Data Analytics". Please refer Unit 2, Section 2.1.

Review Questions

Here are a few ...