**CSE-Data Science/Data Science, IV semester**

**CD404 INTRODUCTION TO DATA SCIENCE**

**UNIT 1**

**INTRODUCTION TO DATA SCIENCE**

**NOTES**

**SYLLABUS**

Unit – I: Introduction

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

Unit – II: Data Collection and Data Pre-Processing Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

Unit – III: Exploratory Data Analytics Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA.

Unit – IV: Model Development Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

Unit – V: Model Evaluation Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

CONTENTS

DATA

SCIENCE

DATA SCIENCE

1. **Data**

In a general sense, "data" refers to raw facts, information, or observations that can be collected, stored, and processed. Data can be in various forms, and it is the basic building block of information. In the context of computing and technology, data is often represented in a structured format that can be manipulated and analysed by computers. Data can be processed and analysed to extract meaningful insights, patterns, and trends. In the context of data science and technology, the term "data" often refers to information that is stored and manipulated electronically, such as in databases, spreadsheets, or other digital formats. This electronic data is crucial for various applications, including analytics, machine learning, and decision-making processes.

Data is a collection of information gathered by observations, measurements, research or analysis. They may consist of facts, numbers, names, figures or even description of things. Data is organized in the form of graphs, charts or tables. Data refers to a systematic record of a specific quantity. In other words, it is a set of facts and figures which are useful in a particular purpose like a survey or an analysis.



Fig 1.1 Data

Data refers to raw facts, figures, or information, often in the form of numbers, text, images, or other forms of input. It can be structured or unstructured and may come from various sources such as observations, measurements, surveys, or databases. Data becomes meaningful when it is processed, analyzed, and interpreted to derive insights, make decisions, or solve problems. In today's digital age, data is increasingly valuable as organizations rely on it to understand trends, make predictions, optimize processes, and enhance decision-making.

* 1. **Types of Data**

Structured Data: - Structured data refers to data that is organized and formatted in a way that is easily understandable and processed by computers. It is typically organized into rows and columns, similar to a table, where each row represents a single record or observation, and each column represents a specific attribute or variable. Structured data follows a predefined schema or data model, which specifies the type and format of each attribute and the relationships between different pieces of data.

e.g. Customer information in a database: This could include fields such as customer ID, name, address, email, phone number, and purchase history. Each customer record would have the same set of attributes arranged in a consistent format.

Unstructured Data: - Unstructured data refers to data that does not have a predefined data model or organization. Unlike structured data, which is organized into rows and columns or follows a specific schema, unstructured data lacks a consistent format, making it more challenging to process and analyze using traditional database management systems.

e.g. Multimedia files: Images, videos, audio recordings, and other multimedia content contain rich information but lack a standardized structure. Analyzing multimedia data often involves extracting features or metadata to make sense of the content.

Semi Structured Data:- Semi-structured data lies somewhere between structured and unstructured data. It has some organizational properties, but it doesn't conform to the formal structure of traditional relational databases.

E,g, Web data: Web pages often contain semi-structured data in the form of HTML documents, which include tags to denote different elements such as headings, paragraphs, and links.



Fig 1.2 Types of Data

* 1. **Importance of Data**

Data is the lifeblood of data science, serving as the foundation upon which insights, predictions, and decisions are built. Its importance in the field of data science cannot be overstated, as it underpins every step of the data science process, from data collection and preprocessing to analysis, modeling, and interpretation.

1. **Information**

Information is a collection of data that gives everyone an advantage since it helps in decision-making for the individuals involved. Data is the term used to describe information entered into and saved on a computer. Information is a fundamental concept in human understanding, communication, and technology. It plays a central role in our daily lives, shaping how we understand the world, make decisions, and interact with others.

At its core, information can be defined as knowledge communicated or received concerning a particular fact or circumstance. It is the result of processing, organizing, and presenting data in a meaningful context. Unlike data, which consists of raw facts or observations, information carries meaning and relevance. It provides insights, answers questions, and helps us make sense of the world around us.

1. **Difference between data and information**

| **S.NO** | **DATA** | **INFORMATION** |
| --- | --- | --- |
| 1 | Data are the variables that help to develop ideas/conclusions. | Information is meaningful data. |
| 2 | Data are text and numerical values. | Information is refined form of actual data. |
| 3 | Data doesn’t rely on Information. | While Information relies on Data. |
| 4 | Bits and Bytes are the measuring unit of data. | Information is measured in meaningful units like time, quantity, etc. |
| 5 | Data can be easily structured as the following: 1.Tabular data 2.Graph 3.Data tree | Information can also be structured as the following: 1. Language 2. Ideas 3. Thoughts |
| 6 | Data does not have any specific purpose | Information carries a meaning that has been assigned by interpreting data. |
| 7 | It is low-level knowledge. | It is the second level of knowledge. |
| 8 | Data does not directly helps in decision making. | Information directly helps in decision making. |
| 9 | Data is collection of facts, which it self have no meaning. | Information puts those facts into context. |
| 10 | Example of data is student test score. | Example of information is average score of class that is derived from given data. |

1. **What is Data Science ?**

Data Science is a field or domain which includes and involves working with a huge amount of data and using it for building predictive and prescriptive analytical models. It’s about capturing, (building the model) analysing (validating the model), and utilizing the data (deploying the best model). It is an intersection of Data and computing. It is a blend of the field of Computer Science, Business, and Statistics together.

Data Science is most revolutionary technology of the Era. It’s all about deriving hidden insights from the data in order to solve real world complex Problems. Data science is one of the most trending and end demanding technologies right now.Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It combines expertise from various domains such as statistics, computer science, mathematics, and domain-specific knowledge to analyse and interpret complex data sets.

**4.1 Need of Data Science**

* The need for data science arises from the increasing volume, variety, and complexity of data generated in today's digital world.
* Handling Big Data
* Informed Decision-Making
* Predictive Analytics
* Efficiency and Automation
* Risk Management
* Healthcare Improvements
* Optimizing Processes
* Fraud Detection and Security
* Customer Segmentation
* Continuous Improvement
* Scientific Research
* Economic Growth
* Adapting to Change

1. **Big Data**

Big data refers to huge amount of data that cannot be stored and processed using traditional computing approaches with a given time frame. Big data refers to large and complex sets of data that are too massive for traditional methods to handle effectively. Imagine you have a computer. You can store a lot of things on it, like photos, documents, and videos. Now, think of big data as having so much information that it would fill up thousands or even millions of computers.

For example if we try to attach a document that is of 100MB in size to an email we would not be able to do so. As the email system would not support an attachment of this size. Therefore this 100MB of attachment with respect to an email can be referred to as big data.

**5.1 Characteristics of big data**

**Value:** Ultimately, the goal of big data analysis is to extract value and actionable insights from the data. Despite its size and complexity, big data holds the potential to uncover valuable patterns, trends, and correlations that can inform decision-making, drive innovation, and create business value.

**Volume:** Big data involves vast amounts of data generated from various sources, including business transactions, social media interactions, sensor readings, and more. The volume of data can range from terabytes to exabytes and beyond, far exceeding the capacity of conventional databases and storage systems.

**Velocity:** Big data is generated at high speeds and requires real-time or near-real-time processing and analysis. With the proliferation of IoT devices, social media platforms, and online transactions, data streams in rapidly and continuously, posing challenges in terms of capturing, processing, and analyzing the data in a timely manner.

**Variety:** Big data comes in diverse formats and types, including structured, semi-structured, and unstructured data. This includes text documents, multimedia files, sensor data, social media posts, and more. Managing and analyzing such heterogeneous data requires flexible and scalable tools and techniques.

**Veracity:** Big data may suffer from inaccuracies, inconsistencies, or biases, posing challenges in ensuring data quality and reliability. Veracity refers to the trustworthiness and reliability of the data, which is crucial for making informed decisions and deriving meaningful insights.

**5.2 Big Data Analytics**

Big data Analytics is the process to extract meaningful insights from big such as hidden patterns, unknown correlations, market trends and customer preferences.

**5.3 Sources Big Data**

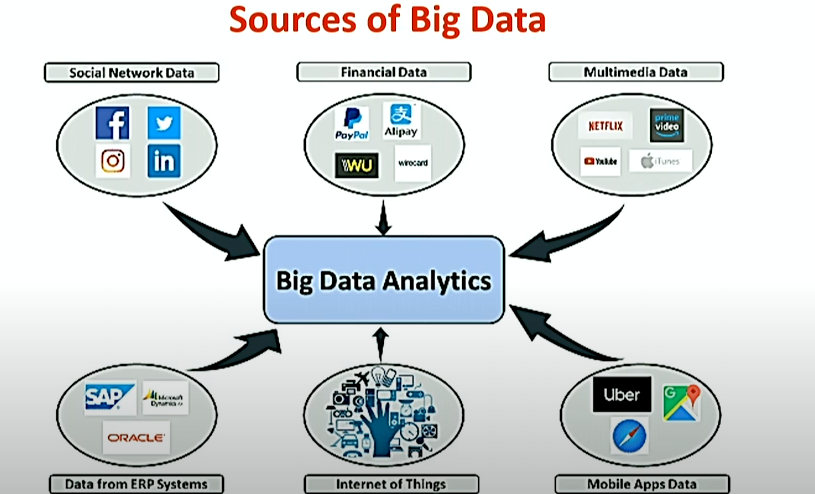


Fig 1.3 Sources Big Data

Social Media: Platforms like Facebook, Twitter, Instagram, and LinkedIn generate vast amounts of data through user interactions, posts, comments, likes, shares, and more. This data provides valuable insights into user behavior, preferences, trends, and sentiments.

Financial Transactions: Banks, financial institutions, and payment processors generate massive amounts of data through transactions, account activity, credit card transactions, stock market trades, and more. Analyzing financial data can help detect fraud, identify market trends, and make investment decisions.

Multimedia and Entertainment: Streaming platforms, video-sharing websites, gaming platforms, and digital media services generate vast amounts of multimedia data, including videos, music, images, and text. Analyzing multimedia data can improve content recommendations, user engagement, and personalized experiences.

Web and E-commerce: Online platforms, websites, and e-commerce stores generate large volumes of data related to user interactions, browsing behavior, purchases, transactions, and clickstream data. Analyzing this data can help businesses understand customer preferences, improve website usability, and optimize marketing strategies.

Internet of Things (IoT) Devices: IoT devices, such as sensors, wearables, smart appliances, and industrial equipment, generate streams of data related to environmental conditions, machine performance, user behavior, and more. This data can be used for predictive maintenance, remote monitoring, and optimization of processes.

**5.4 Big Data Technologies**

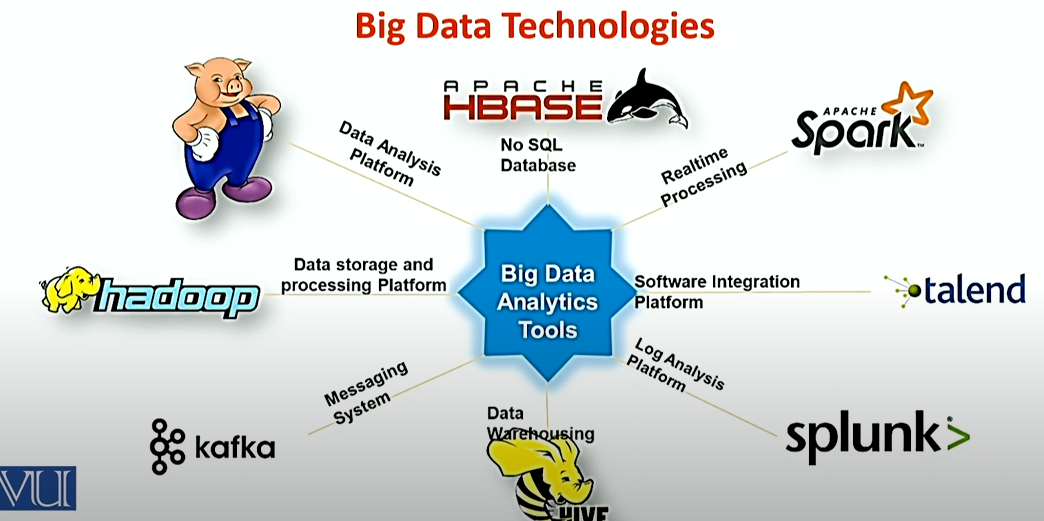


Fig 1.4 Big Data Technologies

Apache Hadoop: Hadoop is an open-source framework for distributed storage and processing of large datasets across clusters of commodity hardware. It consists of two main components: Hadoop Distributed File System (HDFS) for storing data and MapReduce for processing data in parallel across multiple nodes. Hadoop is well-suited for batch processing of structured and unstructured data.

Apache Spark: Spark is a fast and general-purpose distributed computing framework that extends the MapReduce model to support more sophisticated processing tasks, including batch processing, stream processing, machine learning, and graph processing. Spark's in-memory computing capabilities make it faster than Hadoop for iterative algorithms and interactive analytics.

Apache Kafka: Kafka is a distributed streaming platform used for building real-time data pipelines and streaming applications. It provides high-throughput, fault-tolerant messaging between producers and consumers and supports features like message partitioning, replication, and retention. Kafka is commonly used for real-time event processing, log aggregation, and data integration.

NoSQL Databases: NoSQL (Not Only SQL) databases are designed to handle large volumes of unstructured and semi-structured data more flexibly than traditional relational databases. Common types of NoSQL databases include document databases (e.g., MongoDB), key-value stores (e.g., Redis), column-family stores (e.g., Apache Cassandra), and graph databases (e.g., Neo4j). NoSQL databases are used for storing and querying diverse data types at scale.

Apache Hive: Hive is a data warehouse infrastructure built on top of Hadoop for querying and analyzing large datasets using a SQL-like query language called HiveQL. It enables data analysts and data scientists to perform ad-hoc queries, data summarization, and data analysis tasks using familiar SQL syntax. Hive is commonly used for batch processing and data warehousing in Hadoop environments.

**5.5 Data Lake**

A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data. It can store data in its specific format and process any variety of it, ignoring size limits.

**5.6 Big Data and its Challenges**

It is huge, large, or voluminous data, information, or the relevant statistics acquired by large organizations. Many software and data storages is created and prepared as it is difficult to compute the big data manually. It is used to discover patterns and trends and make decisions related to human behavior and interaction technology.

**Challenges**

* Requires specialized skills and expertise in data engineering, data management, and big data tools and technologies
* Can be expensive to implement and maintain due to the need for specialized infrastructure and software
* May face privacy and security concerns when handling sensitive data
* Can be challenging to integrate with existing systems and processes

**5.7 Similarities between Big Data and Data Science**

* Both fields deal with large amounts of data and require specialized skills and expertise
* Both aim to extract insights and knowledge from data to inform decision-making
* Both have a wide range of applications in various industries
* Both can lead to significant cost savings and operational efficiencies when applied correctly
  1. **Differences between Big Data and Data Science**

|  |  |
| --- | --- |
| **Data Science** | **Big Data** |
| Data Science is an area. | Big Data is a technique to collect, maintain and process huge information. |
| It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual. | It is about extracting vital and valuable information from a huge amount of data. |
| It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics. | It is a technique for tracking and discovering trends in complex data sets. |
| The goal is to build data-dominant products for a venture. | The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects. |
| Tools mainly used in Data Science include SAS, R, Python, etc | Tools mostly used in Big Data include Hadoop, Spark, Flink, etc. |
| It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques. | It is a sub-set of Data Science as mining activities which is in a pipeline of Data science. |
| It is mainly used for scientific purposes. | It is mainly used for business purposes and customer satisfaction. |

1. **Data Scientist , Data Analyst and Data Engineer**

**Data Scientists**

* A data scientist is an analytics professional who is responsible for collecting, analyzing and interpreting data to help drive decision-making in an organization.
* A data scientist is a professional who uses a combination of skills in statistics, mathematics, programming, and domain-specific knowledge to analyze and interpret complex data sets.
* The role of a data scientist involves extracting meaningful insights and knowledge from data, helping organizations make informed decisions and solve complex problems.

**What does data Scientist do?**

Data scientists play a crucial role in extracting meaningful insights and knowledge from large and complex datasets. Their responsibilities encompass a wide range of tasks, and they typically engage in the following activities:

* Problem Definition
* Data Collection
* Data Cleaning and Pre-processing
* Exploratory Data Analysis (EDA)
* Feature Engineering
* Model Building
* Interpretation and Insights
* Data Visualization
* Continuous Learning and Improvement
* Deployment and Integration
* Monitoring and Maintenance

Overall, data scientists use their expertise to turn raw data into actionable insights, drive decision-making processes, and contribute to the success and innovation of organizations across various industries.

**6.1 Data Scientist, Data Analyst and Data Engineer**

**Data Scientist**

Designing and development of new processes for data modelling and production using prototypes, algorithms, predictive models and custom analysis. A data scientist analyses and interpret complex data. They are data wranglers who organize (big) data. A data scientist is an analytics professional who is responsible for collecting, analyzing and interpreting data to help drive decision-making in an organization.

The data scientist role combines elements of several traditional and technical jobs, including mathematician, scientist, statistician and computer programmer.

It involves the use of advanced analytics techniques, such as machine learning and predictive modelling, along with the application of scientific principles.

**Roles of Data Science**

* Data science encompasses a diverse set of roles, each playing a crucial part in the end-to-end process of extracting insights from data.
* At the core is the Data Scientist, responsible for conducting Exploratory Data Analysis (EDA), building predictive models, and collaborating with domain experts to translate data into actionable insights.
* Data Analysts focus on cleaning, transforming, and visualizing data, generating descriptive statistics, and supporting decision-making through data interpretation.
* Machine Learning Engineers bring models to life, developing and deploying machine learning algorithms while optimizing for scalability.
* Data Engineers design and maintain data architectures, ensuring data availability, consistency, and reliability through Extract, Transform, Load (ETL) processes.
* Business Intelligence (BI) Analysts leverage data to create informative dashboards and reports, facilitating strategic decision-making within organizations.
* Statisticians contribute their expertise in applying statistical methods to analyze data, conduct experiments, and shape study designs.
* Each role demands a unique skill set, from programming and statistical knowledge to effective communication, highlighting the interdisciplinary nature of data science.
* As the field continues to evolve, the integration of these roles remains pivotal in harnessing the power of data to drive innovation and informed decision-making across industries.

**Core Components of Data Science Roles**

The core components of data science roles encompass a set of fundamental activities that collectively contribute to the process of extracting meaningful insights from data.

* + Data Collection
  + Data Cleaning
  + Exploratory Data Analysis (EDA)
  + Feature Engineering
  + Model Building
  + Model Evaluation
  + Deployment
  + Data Visualization
  + Communication
  + Continuous Learning and Adaptation

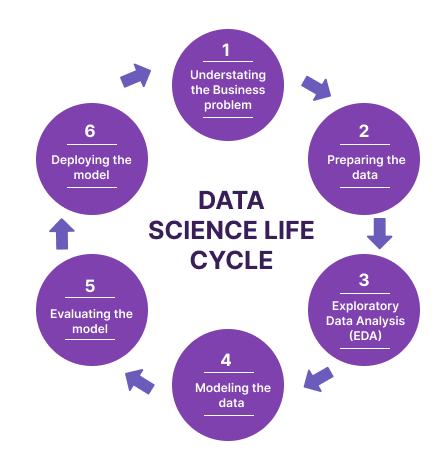


Fig 1.5 Data Science life cycle

**Use Case data scientist:**

* Apply deep learning techniques to process extensive clinical and laboratory reports to conduct a quicker and more precise diagnosis.
* Detect early signs of an issue and enable the doctors to provide preventive care and better treatment to the patients.

**Data Analyst**

Data Analyst analyses numeric data and uses it to help companies make better decisions. A data analyst collects, cleans, and interprets data sets to answer a question or solve a problem.

They work in many industries, including business, finance, criminal justice, science, medicine, and government.



Fig 1.6 Data Analyst Skills

**Role of Data Analyst**

Examine large datasets to identify trends, develop charts, and create visual presentations to help bussiness make more strategic decisions.

**Data Analysis use case:**

Electronic Health Records

* EHRs track and record patient’s health data like pre-existing conditions, allergies and treatments.
* Sharing patients data between healthcare reduces duplication and improve patient care
* Detect early signs of an issue or disease.

**Machine Learning Engineer**

Machine learning engineers act as critical members of the data science team.

Their tasks involve researching, building, and designing the artificial intelligence responsible for machine learning and maintaining and improving existing artificial intelligence systems.

**Data Engineer**

Data Engineer involves in preparing data. They develop, constructs, tests & maintain complete architecture. Data engineers work in various settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret. Their ultimate goal is to make data accessible so that organisations can use it to evaluate and optimise their performance.

**Business Intelligence (BI) Analyst**

A business intelligence (BI) analyst does just that by analyzing complex sets of data within a company to determine recommendations for business growth and improvement.

Knowing how to properly collect and interpret data can have a significant impact on a business' success.

**Statistician**

Statisticians interpret data and communicate results to their clients, often with the aid of mathematical techniques and software.

In this role you'll ensure that complex statistical concepts are explained in a way the client can understand, and advise on strategy.

* 1. **Data Analyst vs Data Engineer vs Data Scientist (Skills)**

**Data Scientist-Skill Set**

* Statistics
* Programming Language
* Data Extraction and processing
* Data wrangling and exploration
* Machine Learning
* Big data processing frameworks
* Data visualization



Fig 1.7 Data Scientist Skill set

Data Science is about combining the extraction of Knowledge from data to answer a particular problem allowing business and stakeholders to make intelligent data driven decisions. Complete technology stack together with soft skills brings more significant challenges to address. Designing and development of new processes for data modelling and production using prototypes, algorithms ,predictive models and custom analysis.

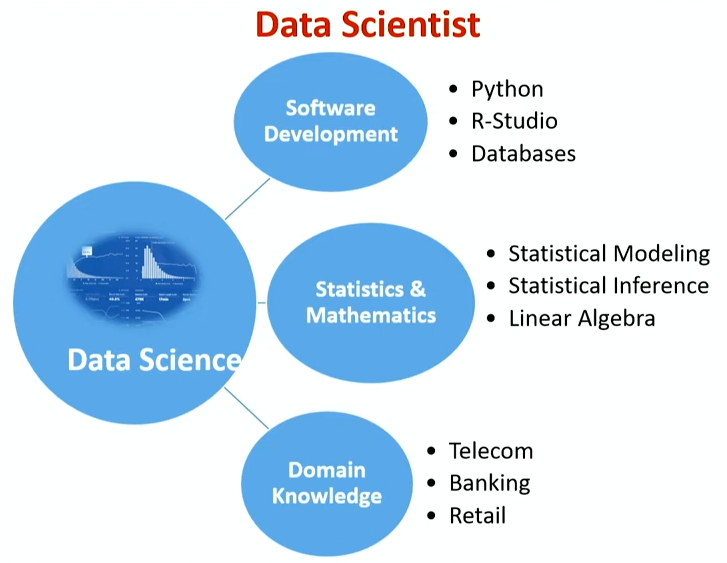


Fig 1.8 Data Scientist Skill set based on parameters

|  |  |  |
| --- | --- | --- |
| **Data Analyst** | **Data Engineer** | **Data Scientist** |
| Data Warehousing | Data Warehousing & ETL(extract transform load) | Statistical & Analytical skills |
| Adobe & Google Analytics | Advanced programming knowledge | Data Mining |
| Programming knowledge | Hadoop-based Analytics | Machine Learning & Deep learning principles |
| Scripting & Statistical skills | In-depth knowledge of SQL/ database | In-depth programming knowledge (SAS/R/ Python coding) |
| Reporting & data visualization | Data architecture & pipelining | Hadoop-based analytics |
| SQL/ database knowledge | Machine learning concept knowledge | Data optimization |
| Spread-Sheet knowledge | Scripting, reporting & data visualization | Decision making and soft skills |

**6.3 Data Analyst vs Data Engineer vs Data Scientist (Roles and Responsibilities)**

|  |  |  |
| --- | --- | --- |
| **Data Analyst** | **Data Engineer** | **Data Scientist** |
| Pre-processing and data gathering | Develop, test & maintain architectures | Responsible for developing Operational Models |
| Emphasis on representing data via reporting and visualization | Understand programming and its complexity | Carry out data analytics and optimization using machine learning & deep learning |
| Responsible for statistical analysis & data interpretation | Deploy ML & statistical models | Involved in strategic planning for data analytics |
| Ensures data acquisition & maintenance | Building pipelines for various ETL operations | Integrate data & perform ad-hoc analysis |
| Optimize Statistical Efficiency & Quality | Ensures data accuracy and flexibility | Fill in the gap between the stakeholders and customer |

**6.4 Organizational Skill Set**

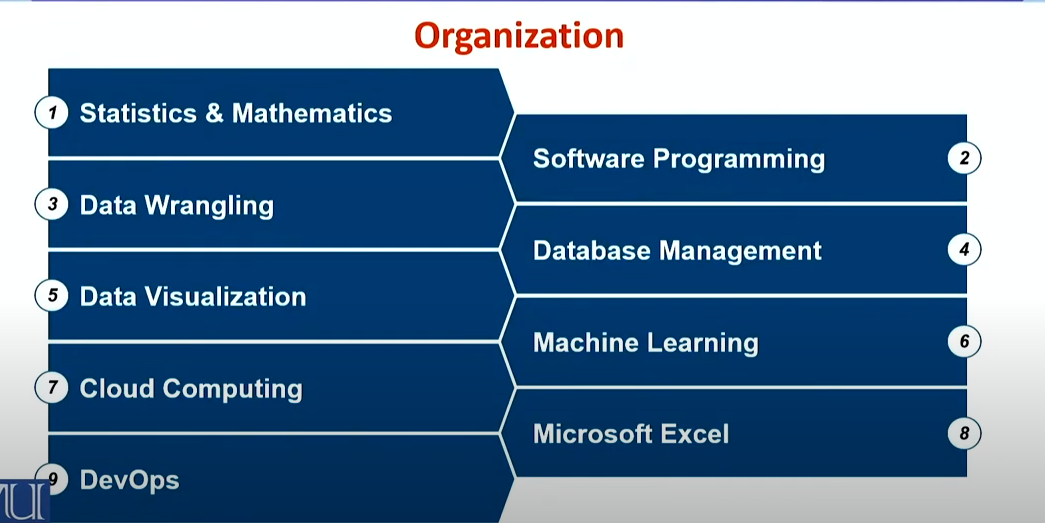


Fig 1.9 Organizational skills

1. **Datafication**

Datafication is the transformation of social actions into online quantified data, thus allowing:

* Real-time tracking
* Data monitoring
* Predictive Analysis and optimization

**Infrastructure Improvement Drive:**

Enterprises are evaluating & implementing infrastructure capacity to cope up with exponential data growth and gain operational maturity in a cost effective manner.

**Data Driven Organization:**

Collective tools, technologies and processes used to transform an organization to a data-driven enterprise. Defining the key to core business operations through a global reliance on data and its related infrastructure.

**Social Media & Datafication**

* Social Media apps usage is playing a great role in datafication:
* Locations visited
* Type of device used
* Type of media viewed
* Interactions with people

**Datafication Based Services**

Datafication is so powerful that it can be used for a nearly infinite number of purposes across multiple of industries.

* Google Maps
* Location Based Services
* Recommendation Engine

**Current Landscape**

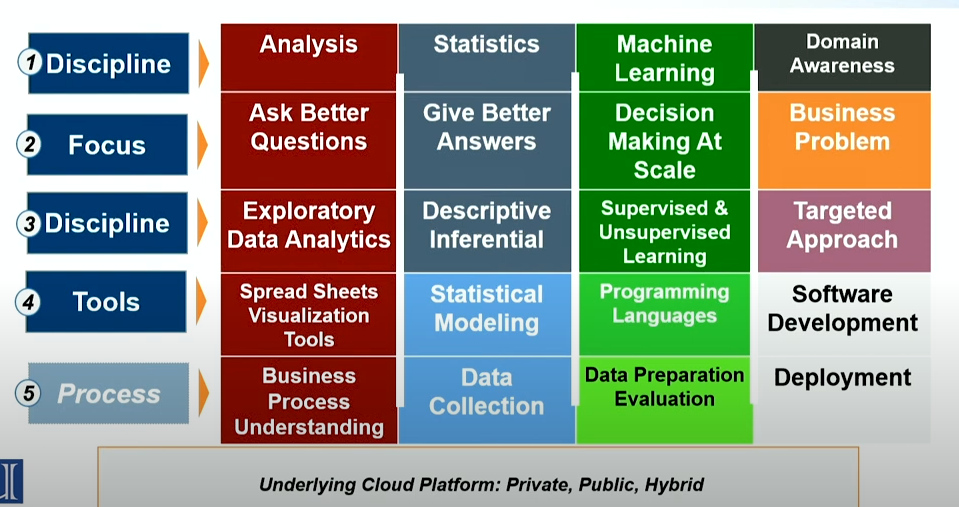


Fig 1.10 Current landscape of data Science project

1. **Software Development**

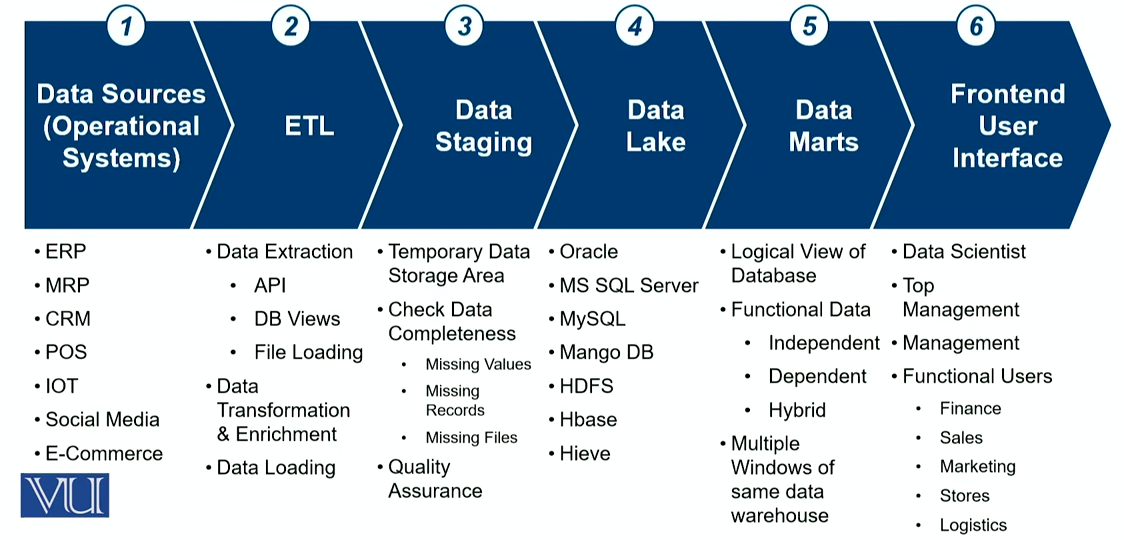


Fig 1.11 Process of software development

**Object Oriented Programming**

Object oriented Programming has an important place in programming skills required for data scientists. Programming languages like Python and Java comply with major OOP principles yet, data scientists need to understand the concepts such as objects , attributes, methods and inheritance) to work in real-world software projects.

**Full Stack Developer**

Data Scientists need to deliver more than machine learning modules in the backend to put the machine learning and data analytics to codes into production following are few examples:

Programming Languages

* Integration Services (Windows, XML, JSON)
* Big data-Hadoop , Mongo DB
* DevOps-Jenkins, Dockers.

**Programming Language**

Data Science has multiple fields but software Engineering is the most important skill in data science.

Following are the most popular data science programming languages:

* + - * + Python
        + R
        + Java
        + C++
        + Scala

**Data Bases SQL & NoSQL**

Data Scientist need to analyse massive data sets and often require SQL and NoSQL tools & techniques.

* + - SQL Query Language
    - HIEVE
    - BASE
    - SPARK

1. **Evolution of Data Science**
2. Early Data Analysis
3. Emergence of Statistics
4. Pre-Computer Era
5. Emergence of Computers and Databases
6. Statistical Methods in the 20th Century
7. Rise of AI & Machine Learning
8. Big Data Era
9. Open Source Tools
10. Cloud Computing
11. Internet of Things (IoT)
12. Artificial Intelligence (AI) and Machine Learning (ML)
13. Evolution of Data Storage
14. Parallel Processing and Distributed Computing
15. Development of Advanced Algorithms
16. Deep Learning and Neural Networks
17. Data Science in Industries

**Early Data Analysis Methods**

Early data analysis methods in data science were characterized by simpler statistical techniques and manual approaches. Traditional data is normally structured data that's organized in records, files, and tables. Fields in traditional data sets are relational, so it's possible to work out their relationship and manipulate the data accordingly.

**Emergence of Statistics**

The emergence of statistics as a formal discipline can be traced back to the 17th century when scholars began to recognize the need for systematic methods to analyse data and draw meaningful inferences. The emergence of statistics was closely linked to the rise of the scientific method and the increasing need to make sense of observational and experimental data. Statistics provided a systematic framework for collecting, organizing, summarizing, analysing, interpreting, and drawing conclusions from data. As the discipline evolved, it became an essential tool in various fields, including biology, economics, sociology, and psychology.

**Pre-Computer Era**

The pre-computer era in data science refers to a time when data analysis and statistical methods were primarily conducted through manual and mathematical processes. During the pre-computer era, data analysis relied heavily on manual methods, where statisticians and analysts performed calculations by hand or using simple mechanical devices. Techniques such as manual sampling, graphical representation of data through charts and graphs, and basic statistical methods like mean and median calculations were employed. The analysis of data was often constrained by the limitations of manual computation, making it challenging to handle large datasets or complex statistical models.

**Development of Computers and Databases**The development of computers and databases marked a transformative period in the history of data science, providing the tools and infrastructure necessary for the systematic storage, retrieval, and analysis of large volumes of data.

Early Computers and Punch Card

* + - Database Management Systems (DBMS)
    - Relational Databases
    - Evolution of Storage Media
    - Integration with Data Science

**Statistical Methods in the 20th Century**

The 20th century was a period of significant advancement in statistical methods, with key developments shaping the way data was analysed and interpreted. Statisticians during this time built upon the foundations laid in previous centuries and made substantial contributions to both theoretical and applied statistics.

**Rise of AI & Machine Learning**

The emergence of machine learning (ML) represents a significant paradigm shift in the field of data science, transforming the way computers learn from data to make predictions or decisions without explicit programming. While the roots of machine learning can be traced back to early concepts, its formalization and widespread adoption began to take shape in the mid-20th century.

**Big Data ERA**

The term big data has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

**Open Source Tools**

Open source software (OSS) is software that is distributed with its source code, making it available for use, modification, and distribution with its original rights.

**Cloud Computing**

Cloud computing is the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the internet (“the cloud”) to offer faster innovation, flexible resources.

Data sciences involves analysing data, while cloud computing involves storing and processing data.

**Internet of Things**

The Internet of Things (IoT) describes the network of physical objects “things” that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

**Artificial Intelligence & Machine Learning**

Artificial intelligence refers to the general ability of computers to emulate human thought and perform tasks in real-world environments, while machine learning refers to the technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience.

**Evolution of Data Storage**

The evolution of storage technologies in data centers has been remarkable, from the first HDDs in the 1950s to the cutting-edge SSDs and hybrid arrays of today.

As data centers continue to play a vital role in our digital age, businesses need reliable and efficient solutions to stay ahead.

**Parallel processing and distributed computing**

In parallel computing, all processors share the same memory and the processors communicate with each other with the help of this shared memory.

Distributed computing systems, on the other hand, have their own memory and processors.

**Deep Learning & Neural Networks**

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.

It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

**Data Science in Industries**

Improving Production and Distribution Efficiencies - The use of data science tools like ML and predictive algorithms can speed up production time, avoid processing delays, and schedule preventative maintenance.

1. **Trends Shaping Data Science**

Several trends were shaping the field of data science. Keep in mind that the field evolves rapidly, and newer trends may have emerged.

* AI and Machine Learning Integration
* Explainable AI (XAI)
* Automated Machine Learning (AutoML)
* Natural Language Processing (NLP) Advancements
* Edge Computing for Data Processing
* Responsible AI

**Advancements in Technology**

Advancements in technology have significantly propelled the field of data science, revolutionizing the way data is collected, processed, and analyzed.

The advent of high-performance computing systems, cloud infrastructure, and distributed computing frameworks has empowered data scientists to tackle complex problems on an remarkable scale.

The evolution of big data technologies, such as Apache Hadoop and Apache Spark, has enabled the efficient processing of massive datasets, facilitating insights into patterns and trends that were previously challenging to discern.

**Impact on Industries**

The evolution of data science has had a profound impact on various industries, transforming the way organizations operate, make decisions, and derive insights from data.

* Healthcare
* Finance
* Retail and E-Commerce

The impact of data science on industries is pervasive, fostering innovation, efficiency, and strategic decision-making.

As organizations continue to leverage data science, the potential for transformative advancements across diverse sectors remains significant.

The evolving landscape of data science continues to shape the way industries operate and adapt in an increasingly data-driven world.

**Future Directions**

The future directions of data science hold exciting prospects as the field continues to evolve with technological advancements and societal needs.

Explainable AI (XAI) is likely to gain prominence, addressing the interpretability of complex machine learning models and ensuring transparency in decision-making.

The integration of data science with emerging technologies, such as quantum computing, may unlock new possibilities for solving computationally intensive problems.

Continued advancements in natural language processing (NLP) and language models could lead to more sophisticated and context-aware AI systems, impacting areas like conversational AI and language translation.

1. **Process of data science/Stages in data science project**

Data science is widely used in various industries, including finance, healthcare, marketing, and technology, to inform decision-making processes, predict future trends, and gain a competitive edge. It involves a combination of technical skills, domain knowledge, and critical thinking to extract meaningful information from data.

**Business Understanding**

* Business Understanding: This phase involves understanding the project's objectives, defining the business problem, and determining the success criteria.
* Data Understanding: In this phase, the relevant data sources are identified, data quality is assessed, and initial insights are gained.
* The data scientists in the room are the people who keep asking the why’s.
* They’re the people who want to ensure that every decision made in the company is supported by concrete data, and that it is guaranteed (with a high probability) to achieve results.
* Before you can even start on a data science project, it is critical that you understand the problem you are trying to solve.
* We typically use data science to answer five types of questions:
  + - * How much or how many? (regression)
      * Which category? (classification)
      * Which group? (clustering)
      * Is this weird? (anomaly detection)
      * Which option should be taken? (recommendation)
      * In this stage, you should also be identifying the central objectives of your project by identifying the variables that need to be predicted.

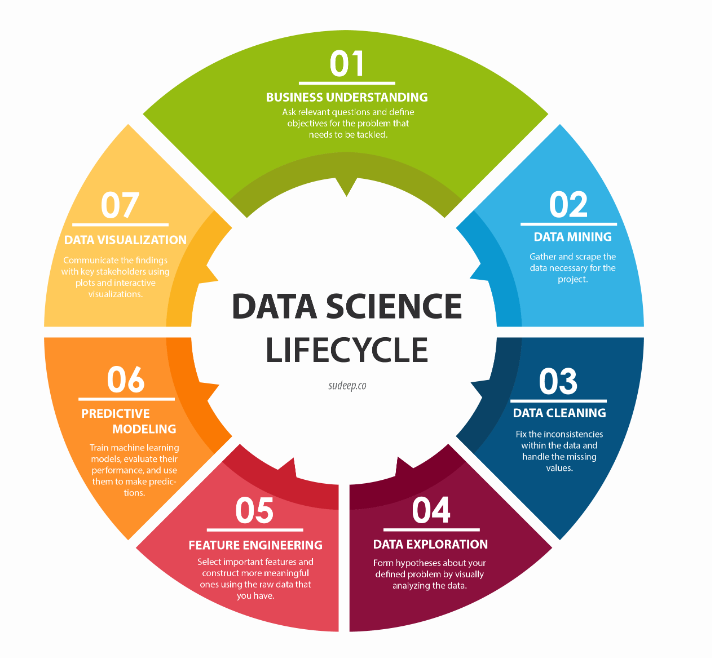


Fig 1.12 Data Science Lifecycle

**Data Mining**

* Data mining is the process of gathering your data from different sources.
* At this stage, some of the questions worth considering are –
  + What data do I need for my project?
  + Where does it live?
  + How can I obtain it?
  + What is the most efficient way to store and access all of it?

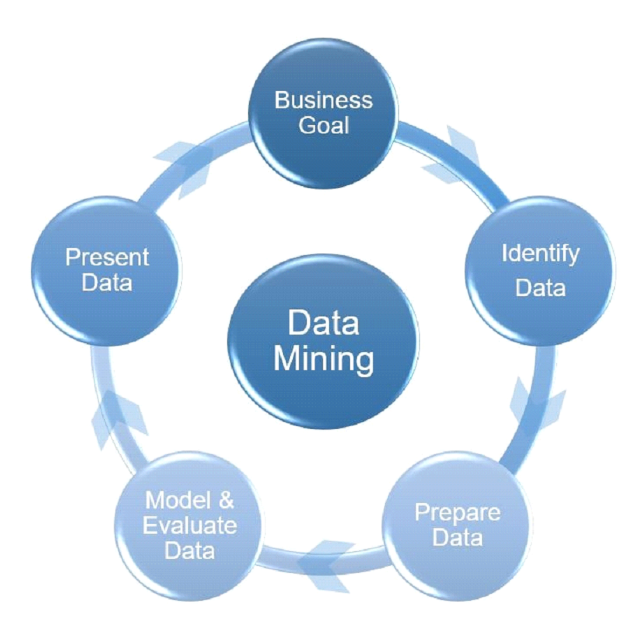


Fig 1.13 Data Mining

**Data Collection:** Gathering relevant data from various sources, which can be structured (organized in a tabular format) or unstructured (text, images, videos)

**Data Acquisition:-**Now you found the objectives of the project now its time to start gathering the data.so data mining is the process of gathering your data from different sources.

**Data Cleaning and Pre-processing:** Cleaning and organizing the data to remove errors, missing values, and inconsistencies. This step is crucial for ensuring the quality and reliability of the analysis. Transform data into desired format. Removing Missing values, Corrupted data, Remove unnecessary data. We move on to the most time-consuming step of all - cleaning and preparing the data. This is especially true in big data projects, which often involve terabytes of data to work with.The reason why this is such a time consuming process is simply because there are so many possible scenarios that could necessitate cleaning.

* For instance, the data could also have inconsistencies within the same column, meaning that some rows could be labelled 0 or 1, and others could be labelled no or yes. The data types could also be inconsistent - some of the 0s might integers, whereas some of them could be strings.
* One of the steps that is often forgotten in this stage, causing a lot of problems later on, is the presence of missing data.
* Missing data can throw a lot of errors in the model creation and training.

**Exploratory Data Analysis (EDA):** Exploring and summarizing the main characteristics of the data using statistical and visualization techniques to gain a better understanding of its patterns and trends.

* Understand the patterns in the data
* Retrieve useful insight
* Form hypotheses

So now that you have sparkling clean set of data, you are finally ready to get started with your analysis.

The data exploration stage is basically the brainstorming of data analysis so in order to understand the patterns in your data you can use histogram or you can even create interactive visualizations. This time you try to explore the different models that can be applied your data.The data exploration stage is like the brainstorming of data analysis. This is where you understand the patterns and bias in your data.

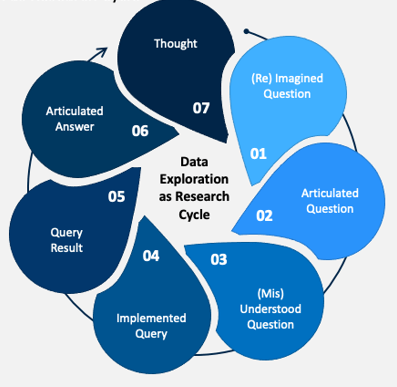


Fig 1.14 Data Exploration

It could involve pulling up and analyzing a random subset of the data using Pandas, plotting a histogram or distribution curve to see the general trend, or even creating an interactive visualization that lets you dive down into each data point and explore the story behind the outliers.

**Feature Engineering:** Creating new features or transforming existing ones to improve the performance of machine learning models. In machine learning, a feature is a measurable property or attribute of a phenomenon being observed. If we were predicting the scores of a student, a possible feature is the amount of marks they get.

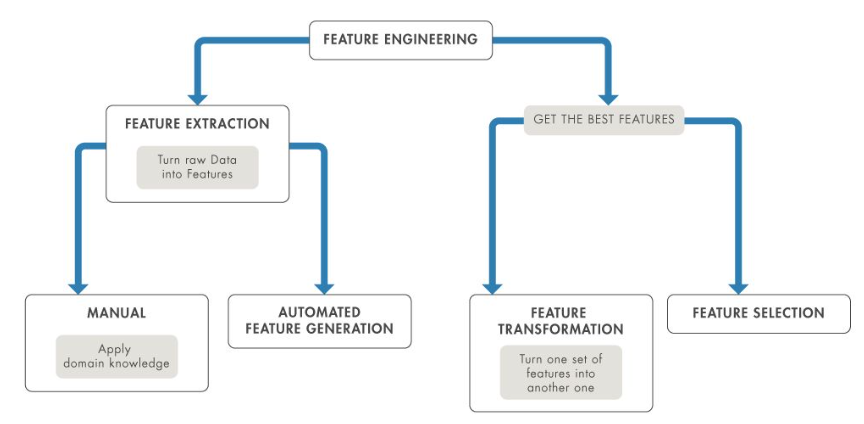


Fig 1.15 Feature Engineering

Feature engineering is the process of using domain knowledge to transform your raw data into informative features that represent the business problem you are trying to solve. This stage will directly influence the accuracy of the predictive model you construct in the next stage. Feature selection is the process of cutting down the features that add more noise than information.

**Model Building and Machine Learning:** Applying various machine learning algorithms to build predictive models or uncover patterns and relationships within the data. Determine optimal data features for the machine learning model. Create a model that predicts the target most accurate. Evaluate and test the efficiency of the model. Predictive modeling is where the machine learning finally comes into your data science project. Predictive modeling is a commonly used statistical technique to predict future behaviour. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

**Data Visualization**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.



Fig 1.16 Data Visualization

**Model Evaluation and Validation:** Assessing the performance of the models using appropriate metrics and ensuring their reliability on new, unseen data.

Interpretation and Visualization: Communicating the results of the analysis to non-technical stakeholders through visualizations and reports, and deriving actionable insights.

**Deployment:**

* Check the deployment environment for dependency issue.
* Deploy the model in a pre-production/test environment
* Monitor the performance.

A goal of this stage is to deploy the model into a production or maybe a production like environment. This is basically done for final user acceptance and the users have to validate the performance of the models and if there are any issues with the model or any issue with the algorithms then they have to be fixed in this stage.

1. **Applications of Data Science**

Data science has a wide range of applications across various fields, helping organizations make informed decisions, optimize processes, and gain valuable insights.

**Healthcare**

* Predictive analytics for patient outcomes.
* Disease diagnosis and early detection.
* Drug discovery and development.
* Personalized medicine based on patient data.

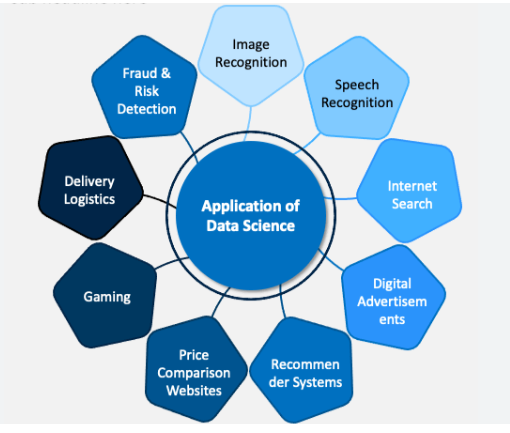


Fig 1.17 Applications of Data Science

**Finance**

* Fraud detection and prevention.
* Credit scoring and risk assessment.
* Algorithmic trading and financial modeling.
* Customer segmentation for targeted marketing.

**Retail**

* Demand forecasting and inventory optimization.
* Customer segmentation and personalized marketing.
* Price optimization and dynamic pricing.
* Recommender systems for product recommendations.

**Education**

* Personalized learning platforms.
* Predictive analytics for student performance.
* Admissions and enrollment optimization.
* Educational resource allocation.

**Transportation and Logistics**

* Route optimization and fleet management.
* Demand forecasting for transportation services.
* Predictive maintenance for vehicles and equipment.
* Supply chain optimization.

**Marketing and Advertising**

* Customer segmentation and targeting.
* Campaign optimization and A/B testing.
* Social media analytics.
* Customer sentiment analysis.

**Human Resources**

* Employee performance analytics.
* Recruitment and talent acquisition optimization.
* Employee satisfaction and engagement analysis.
* Workforce planning and optimization.

**Environmental Science**

* Climate modeling and prediction.
* Environmental monitoring and analysis.
* Wildlife conservation and biodiversity studies.
* Pollution tracking and control.

**Image Processing**

* Data science techniques, particularly deep learning algorithms, are used for image classification tasks. Convolutional Neural Networks (CNNs) are commonly employed to classify images into predefined categories, such as identifying objects in photographs or detecting specific patterns.
* Data science enables the development of algorithms for detecting and localizing objects within images. Object detection techniques, such as Faster R-CNN, YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector), utilize data-driven approaches to accurately locate objects and delineate their boundaries within images.
* Data-driven approaches are utilized for enhancing the quality of images and restoring degraded or corrupted images. Techniques such as image denoising, deblurring, and super-resolution leverage data science methodologies to improve image clarity and fidelity.
* Data science plays a crucial role in medical imaging applications, where it aids in the analysis and interpretation of medical images such as X-rays, MRI scans, CT scans, and histopathological images. Machine learning algorithms assist in tasks such as tumor detection, organ segmentation, disease classification, and treatment planning.

**Speech Recognition**

* Data science encompasses the development and training of machine learning models for speech recognition tasks. Supervised learning algorithms, such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and deep learning architectures like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are utilized to recognize speech patterns and predict corresponding text outputs.
* Data science techniques are applied to build language models that capture the statistical properties and structures of natural languages. These models aid in predicting the sequence of words or phonemes given an input speech signal, improving the accuracy of speech recognition systems.
* Data science methods facilitate speaker identification and verification tasks, where the goal is to determine the identity of a speaker based on their voice. Machine learning algorithms are trained to recognize unique characteristics of individual voices, enabling applications such as voice authentication.

**Internet search**

* Data science enables search engines to understand the intent behind user queries by analyzing the context, semantics, and user behavior associated with search queries. Machine learning algorithms are employed to classify and categorize queries, extract user intent, and personalize search results based on user preferences and past interactions.
* Data science techniques are utilized in web crawling and indexing processes, which involve systematically browsing the web to discover and index new web pages or content. Crawling algorithms employ strategies such as breadth-first search, depth-first search, and heuristics to efficiently traverse the web and collect relevant data for indexing and retrieval.
* Data science enables search engines to personalize search results for individual users based on their search history, demographic information, geographic location, and browsing behavior. Personalization algorithms utilize collaborative filtering, content-based filtering, and recommendation systems to tailor search results to the preferences and interests of each user.
* Data science methods are used to improve search accuracy by expanding and correcting user queries to account for variations in spelling, synonyms, and related terms. Techniques such as query expansion, spell correction, and synonym detection help refine user queries and enhance the retrieval of relevant information.

**Recommendation system**

* Data science techniques are used to create user profiles and segment users based on their behavior, demographics, and preferences. Clustering algorithms and dimensionality reduction techniques help identify similar groups of users, enabling personalized recommendations tailored to each user segment.
* Data science facilitates content-based filtering methods, which recommend items to users based on the similarity between items and user preferences. Natural language processing (NLP) and machine learning algorithms are employed to analyze item attributes, such as text descriptions or features, and generate item representations for recommendation.
* Data science enables the development of hybrid recommendation systems that combine multiple recommendation approaches, such as collaborative filtering, content-based filtering, and hybrid models. Ensemble methods and machine learning techniques are used to integrate diverse recommendation strategies and optimize recommendation performance.

**Price comparison website**

* Data science facilitates product matching and categorization tasks, where similar products from different retailers or brands need to be identified and grouped together. Machine learning algorithms, text mining techniques, and product embeddings help match products based on their attributes, descriptions, and features.
* Data science enables price tracking and historical analysis functionalities, allowing users to monitor price fluctuations over time and make informed purchasing decisions. Time series analysis, trend detection algorithms, and anomaly detection methods help identify price trends, discounts, and seasonal variations in product prices.
* Data science techniques are utilized to enhance search and filtering capabilities on price comparison websites, enabling users to efficiently find products based on their preferences, specifications, and price ranges. Natural language processing (NLP) algorithms, faceted search, and recommendation systems help improve search relevance and user experience.

**Gaming**

* Data science techniques are used to analyze player behavior, interactions, and preferences within games. Game developers employ data mining, machine learning, and statistical analysis to understand how players navigate through game levels, make decisions, and engage with game content. This insight informs game design decisions, such as level difficulty adjustments, content updates, and personalized gameplay experiences.
* Data science helps predict player retention and identify churn risks by analyzing player activity patterns and engagement metrics. Predictive modeling techniques, such as survival analysis and machine learning algorithms, are employed to forecast player churn and implement targeted retention strategies, such as personalized offers, rewards, and content recommendations to keep players engaged and loyal.

**Fraud Detection**

* Data science techniques are used to detect anomalies or unusual patterns in data that may indicate fraudulent activity. Anomaly detection algorithms, such as statistical methods, clustering techniques, and machine learning models like Isolation Forests and Autoencoders, help identify transactions, behaviors, or events that deviate significantly from the norm.
* Data science enables the recognition of patterns associated with fraudulent behavior by analyzing historical data and identifying common fraud schemes or tactics. Machine learning algorithms, including supervised learning and unsupervised learning techniques such as association rule mining and sequence pattern mining, help uncover hidden patterns and relationships in data that may indicate fraudulent activity.
* Data science enables the analysis of behavioral patterns and trends associated with legitimate users and fraudulent actors. Machine learning algorithms, such as clustering and classification models, help differentiate between normal and abnormal behavior by analyzing user interactions, transaction histories, and usage patterns, enabling proactive detection of suspicious activity.

1. **Advantages of Data Science**

### **Better Decision Making**

Data Science allows organizations to make informed decisions based on data and evidence. By analyzing vast amounts of data, **Data Scientists** can identify patterns and trends and provide insights that can inform decision-making processes.

### **Improved Customer Experience**

Data Science can help organizations better understand their customers by analyzing customer data. This can lead to more personalized and targeted marketing campaigns, better product development, and an overall improved customer experience.

### **Increased Efficiency**

Data Science can automate repetitive and time-consuming tasks, freeing up employees to focus on higher-value activities. It can also identify inefficiencies in processes and systems, allowing organizations to make improvements and increase efficiency.

### **Increased Revenue**

By analyzing sales data and identifying trends and patterns, Data Science can help organizations to increase their revenue. For example, it can identify the most profitable products or target markets, allowing organizations to focus their resources on these areas.

### **Better Fraud Detection**

Data Science can be used to detect fraud by analyzing large amounts of data to identify anomalies and suspicious activities. This can help organizations to minimize financial losses and protect their customers.

### **Improved Healthcare Outcomes**

Data Science can be used in healthcare to analyze patient data and improve patient outcomes. For example, it can be used to identify patients at high risk of developing certain conditions, allowing healthcare providers to take proactive measures to prevent them.

### **Improved Public Services**

Data Science can be used by government organizations to improve public services. For example, it can be used to analyze crime data to identify hotspots and allocate resources more effectively or to analyze transportation data to improve traffic flow and reduce congestion.

### **Better Environmental Protection**

Data Science can be used to analyze environmental data, allowing organizations to identify and address environmental issues. For example, it can be used to monitor and mitigate climate change impacts or monitor wildlife populations.

1. **Disadvantages of Data Science**

**Technical Complexity:** Data science involves complex technical skills such as coding, statistics, and machine learning, which can be challenging to master.

**Data Quality:** Data scientists need high-quality data to perform accurate analyses, but data quality can be an issue in some cases.

**Time-Consuming:** Data science projects can be time-consuming, requiring several months or even years to complete.

**Limited Resources:** In some cases, data scientists may face limited resources such as hardware, software, or funding

**How to Overcome the Challenges:**

**Continuous Learning:** Data science requires continuous learning and staying up-to-date with the latest tools and techniques. To overcome the technical complexity, data scientists can take online courses or attend workshops to enhance their skills.

**Data Management:** To ensure high-quality data, data scientists should work with domain experts and use data cleaning and preprocessing techniques. This can help to minimize errors and improve the accuracy of the analysis.

**Project Management:** To manage time effectively, data scientists can break down projects into smaller tasks and prioritize them based on their importance. Project management tools such as Trello or Asana can be helpful in managing workflows and deadlines.

**Resource Planning:** To overcome limited resources, data scientists can leverage cloud computing services such as Amazon Web Services (AWS) or Microsoft Azure, which offer scalable and affordable solutions for data storage and computing.

1. **Tools for Data Science**

It is required that they have a clear understanding of the tools that are necessary for the programming to work. We decided to provide a little insight into the tools that can be used for data visualization, statistical programming languages, algorithms, and databases. These tool can be learn using some of the best [Data Science Course](https://www.odinschool.com/datascience-bootcamp). These tools will help speed up your process as you do not have to further search anywhere else for what you need.

* **DataRobot :-** It is a global automated Machine Learning platform. With the capabilities of Data Science, Machine Learning, Statistical Modeling, Artificial Intelligence, Augmented Analytics, Machine Learning Operations (MLOps), Time Series Modeling.
* **MLBASE :-**One of the best Data Science tools and provides distributed and statistical techniques that are key to transforming big data into actionable knowledge. It provides functionality to end-users for a wide variety of standard machine learning tasks such as classification, regression, collaborative filtering, and more general exploratory data analysis techniques
* **Apache Graph :-**Apache Graph supports high-level scalability. It is an iterative graph processing system that has been specially developed for this purpose. This was derived from the Pregel model but comes with more number of features and functionalities when compared with the Pregel model. This open-source model helps data scientists to utilize the underlying potential of structured datasets at a large scale.
* **Apache Spark :-** This is another free tool that offers cluster computing in a blink of the eye, which is at lightning bolt speed. Today, a number of organizations are using Spark for processing large datasets. This data scientist tool is capable of accessing diverse data sources, which include HDFS, HBase, S3, and Cassandra.
* **Cascading :-** It is specifically for data scientists who are building big data apps on Apache Hadoop. It allows users to solve both complex and simple data problems, using cascading. This is because it offers computation engines, data processing, scheduling capabilities, and systems integration framework.
* **TABLEAU :-** It is a Data Science visualization software with powerful graphics to make interactive visualizations. It can interface with databases, spreadsheets, OLAP (Online Analytical Processing) cubes. It provides the capability of visualizing the geographical data and for plotting longitudes and latitudes in maps.
* **TENSORFLOW :-** This is an ML tool, which is widely used for advanced Machine Learning algorithms like Deep Learning. It is an open-source and ever-evolving toolkit which is known for its performance and high computational abilities.
* **SAP HANA :-**It is an effective tool from SAP with SAP HANA Predictive Analysis Library (PAL).
* **MONGODB :-**This is another Data Analysis tool that is quite popular since it allows cross-platform document orientation. It has a basic query and aggregation framework, but to do more advanced analytics. It is a perfect choice to iterate ML training experiments.

1. **Data Science with Cloud Computing and its Security**

**16.1 Data Security Issue**



Fig 1.18 Data Security issue

Data security is a critical concern in today's digital age, as organizations and individuals increasingly rely on technology to store, process, and transmit sensitive information.

Several data security issues pose potential risks to the confidentiality, integrity, and availability of data.

**Unauthorized Access:**

* + Unauthorized users gaining access to sensitive data.
  + Weak or compromised authentication mechanisms, such as passwords.

**Data Breaches:**

* + Unauthorized disclosure or access to confidential information.
  + External attacks, such as hacking, or internal threats from employees.

**Malware and Ransomware:**

* + Malicious software that can compromise data integrity and availability.
  + Ransomware encrypts data and demands payment for decryption.

**Phishing Attacks:**

* + Deceptive attempts to obtain sensitive information by posing as a trustworthy entity.
  + Phishing emails, websites, or messages can trick individuals into revealing passwords or other confidential data.

**Insider Threats:**

* + Intentional or unintentional security risks posed by employees or other trusted individuals.
  + Employees with access to sensitive data may misuse it or accidentally expose it.

**Insecure Data Storage:**

* + Inadequate (insufficient) protection of data at rest, such as unencrypted databases or storage devices.
  + Storing sensitive information on insecure servers or cloud services.

**Inadequate Data Encryption:**

* + Lack of encryption during data transmission or storage.
  + Unencrypted communication channels may be vulnerable to interception.

**Poor Patch Management:**

* + Failure to apply timely security patches and updates to software and systems.
  + Outdated software may have known vulnerabilities that can be exploited

**Physical Security Threats:**

* + Theft or damage to physical devices containing sensitive data.
  + Lack of controls to secure servers, computers, or storage devices.

**Data Residency and Compliance:**

* + Failure to comply (to obey the rules) with data protection regulations and laws.
  + Inadequate (insufficient) measures to ensure data residency requirements are meet.

**Insecure APIs:**

* + Vulnerabilities in application programming interfaces (APIs) that can be exploited.
  + Insecure APIs may expose sensitive data to unauthorized access.

**Lack of User Awareness:**

* + Insufficient education and awareness among users about security best practices.
  + Users may fall victim to social engineering attacks or unintentionally compromise security.

**Data Interception:**

* + Unauthorized interception of data during transmission.
  + Man-in-the-middle attacks can compromise the confidentiality of transmitted information.

**16.2 Data Science with Cloud Computing**

What is Cloud?

Cloud computing is the on-demand availability of computer system,resource,especially:

* + Data storage
  + Computing power
  + RAM
  + Internet Connectivity
  + Without direct active management by the ueser

**Types of Cloud**

Large clouds often have functions distributed over multiple locations, each location being a data centre:

* + Private Cloud
  + Public Cloud
  + Hybrid Cloud

**Cloud Services Models**

Clouds operators often have functions distributed over world:

* + SaaS-Software as a Service
  + DaaS-Data as a Service
  + PaaS- Platform as a Service
  + IaaS-Infrastructure as a Service

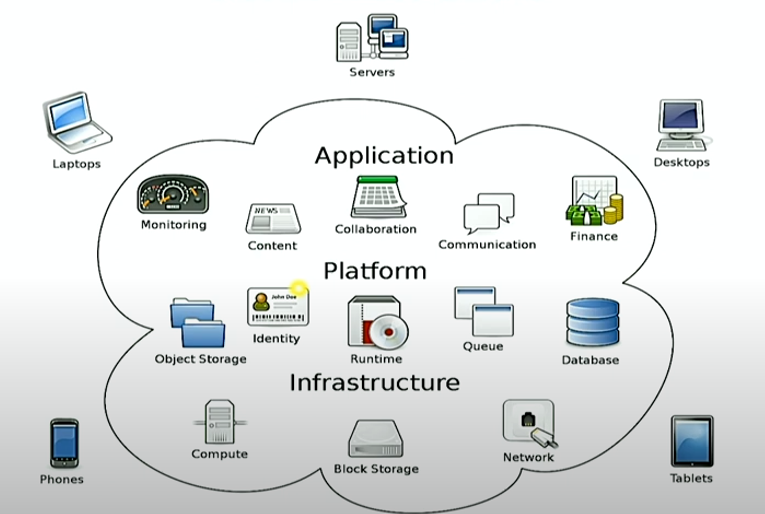


Fig 1.19 Cloud Infrastructure

**Data Science with Cloud**

Data Science and cloud computing essentially go hand in hand. A data Scientist typically analyses different types of data that are stored in the cloud.

**Why It’s Important?**

* + Analysis and storage are the two important challenges for organizations large and small.
  + Amount of big data generated has accelerated tremendously.

**Storage data in Cloud**

* + Economical
  + Secure
  + Cloud Computing Benefits for data Scientist:
  + Use platforms for free or at a minimal price

Choice of available tools:

* + Open Sources (Hadoop,Python,MySQL)
  + Commercial (Oracle, MS SL,SAS,Tableau)

**16.3 Data Science with Cloud: Google Cloud Platforms**

**What is Google Cloud?**

Google Cloud Platform, is a suit of cloud computing services running on the same infrastructure used internally for its end-user products:

* + Google Search
  + Gmail
  + Google Drive
  + YouTube
  + Google Translate

**Google Cloud Benefits:**

* + High Productivity & Innovative
  + Easy Adoptability
  + Remote access
  + Unmatched security
  + Flexibility & Control
  + Allows to Store & Compute Data
  + Manage SDLC
  + API’s Available

**16.4 Data Science with Cloud: Amazon Web Services**

**What is AWS Cloud Platform?**

* + Amazon Web Services, provides on demand cloud computing platforms and APIs to individuals, companies, and governments on a metered pay-as-you-go basis:
  + 100+Products
  + Free Trials + Totally Free
  + Compute,Storage,Databases
  + Management Console
  + Machine learning

**AWS Cloud Platform Benefits**

* + Amazon Web Services, is the primary profit driver for Amzon. It provides:
  + Servers
  + Storage
  + Security
  + Networking
  + Remote Computing
  + Mobile Development

**16.5 Data Science with Cloud: Microsoft Azure Cloud**

**What is MS Azure Cloud Platform?**

* + The azure application service platform runs logical,mobile,API and business applications and automatically manages the resources required by these applications:
  + 200+Products
  + Free Account
  + Hybrid + Multi Cloud
  + Break, Prepare & Optimize
  + Compute, Containers ,Networking

**MS Azure Cloud Platform Benefits:**

* + Cost Optimization
  + Unmatched Speed of Services
  + Global Scale Digital Workplace
  + Productivity & Performance
  + Flexibility, Reliability & Security
  + CICD pipelines(DevOps)
  + Disaster Recovery,Bussiness Continuity
  + Analytics
  + Azure APIs and Workflows

**16.6 Data Science with Cloud: Social Media Analytics on Cloud**

**Social Media Analytics on Cloud**

* + SMAC is an ecosystem enabling business transformation from e-business to digital and improve business operations and with minimal overhead and maximum reach:
  + Social Media
  + Mobile Applications
  + Data Analysis
  + Cloud Computing

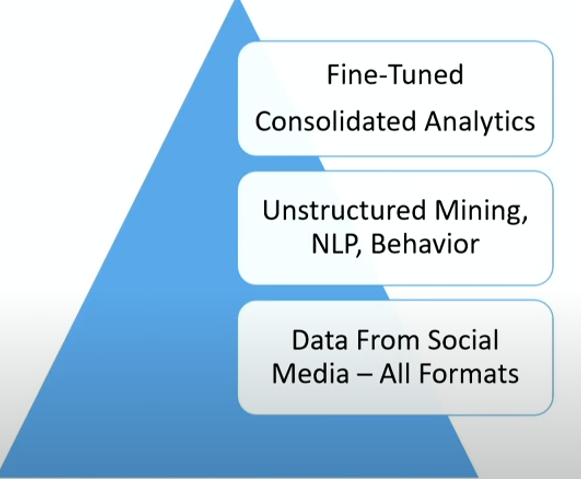


Fig 1.20 Social media Analytics on Cloud

**Social Media Analytics on Cloud offers the following benefits:**

* + Integrated Data Storage
  + Cost Effective
  + Robust & Fast Computations
  + Multi-Tasking
  + Statistical Modelling
  + AI/ML Algorithms
  + Actionable Insights

VIVA QUESTIONS

1. Define data science and explain its significance in modern society.

Data science is an interdisciplinary field that involves extracting insights and knowledge from large datasets using statistical, computational, and analytical techniques. Its significance in modern society lies in its ability to drive informed decision-making, foster innovation, improve efficiency, personalize experiences, and address complex challenges across various industries, ranging from healthcare and finance to marketing and urban planning.

1. Discuss the interdisciplinary nature of data science and its relationship with fields such as statistics, computer science, and domain expertise.

Data science is inherently interdisciplinary, drawing from fields such as statistics, computer science, and domain expertise. Statistics provides foundational methods for data analysis and inference, while computer science offers tools and techniques for data processing and machine learning. Domain expertise is essential for contextual understanding and interpreting results, ensuring that data-driven insights are relevant and actionable within specific domains such as healthcare, finance, and marketing. This interdisciplinary collaboration enables data scientists to tackle complex problems, innovate solutions, and drive impactful outcomes across diverse domains.

1. What is data science, and why is it important?

Data science is an interdisciplinary field that involves the extraction of insights and knowledge from data using scientific methods, algorithms, and systems. It combines techniques from statistics, computer science, and domain expertise to analyze large volumes of data and make data-driven decisions. Data science is important because it enables organizations to extract valuable insights from data, solve complex problems, and drive innovation and decision-making processes.

1. How has data science evolved over time?

Data science has evolved from traditional statistical methods to modern, data-driven approaches enabled by advancements in technology and computing power. Initially, data analysis primarily involved descriptive statistics and hypothesis testing. With the advent of big data and machine learning, data science has expanded to include predictive modeling, pattern recognition, and deep learning techniques. Today, data science encompasses a wide range of applications, from business analytics to scientific research and artificial intelligence.

1. What are the key roles in a data science team, and what are their responsibilities?

In a data science team, key roles include data scientists, data engineers, and data analysts. Data scientists are responsible for developing and implementing machine learning models and algorithms to extract insights from data. Data engineers focus on building and maintaining data infrastructure, pipelines, and systems for data collection and storage. Data analysts analyze data to generate actionable insights and support decision-making processes.

1. What are the stages involved in a typical data science project?

The stages in a data science project typically include data collection, data preprocessing, exploratory data analysis, modeling, evaluation, and deployment. Data collection involves gathering relevant data from various sources. Data preprocessing involves cleaning, transforming, and preparing the data for analysis. Exploratory data analysis involves exploring the data to identify patterns and insights. Modeling involves building and training machine learning models. Evaluation involves assessing the performance of the models using appropriate metrics. Deployment involves deploying the models into production environments for use.

1. Can you provide examples of data science applications in different fields?

Sure! In healthcare, data science is used for disease prediction, medical image analysis, and personalized medicine. In finance, data science is used for fraud detection, risk management, and algorithmic trading. In marketing, data science is used for customer segmentation, recommendation systems, and predictive analytics. In transportation, data science is used for route optimization, traffic forecasting, and autonomous vehicles.

1. What are some data security issues associated with data science projects?

Data security issues in data science projects include unauthorized access to sensitive data, data breaches, and privacy violations. Inadequate data protection measures, such as weak encryption, insecure data storage, and insufficient access controls, can expose organizations to security risks.