CS442- Introduction of Data Science

Introduction to Data Science

Module# 1

What is Data Science:

Introduction

- How much data YOU produce in day?
- What information is hidden in that data?
- Where that data is going? and
- Who OWNS that data?



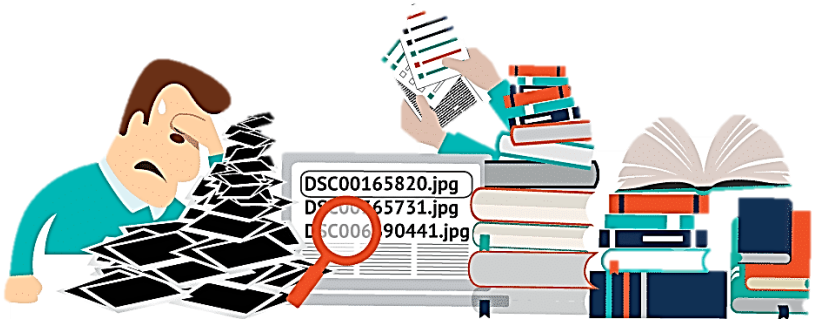
YOU are generating valuable data during all such activities producing valuable information.

Not only for yourself but also for the businesses around the WORLD!




What is Data Science

The process of extracting knowledgeable insights from data using scientific methods and tools.



CS442- Introduction of Data Science

Components of Data Science

There are four major components of data science:

- Data Strategy
- Data Engineering
- Data Modeling & Analysis
- Data Visualization & Operationalization

Data Engineering

End to end data management is called data engineering:

- Understanding data sources (Formats, Types)
- Data Pipelines, Cleansing, Transformation, Enrichment
- Data Retention Management
- Data Marts
- Data Analysis & Visualization

Data Modeling & Analysis

- Data science is combination of various statistical models and software development techniques.

Data Visualization & Operationalization

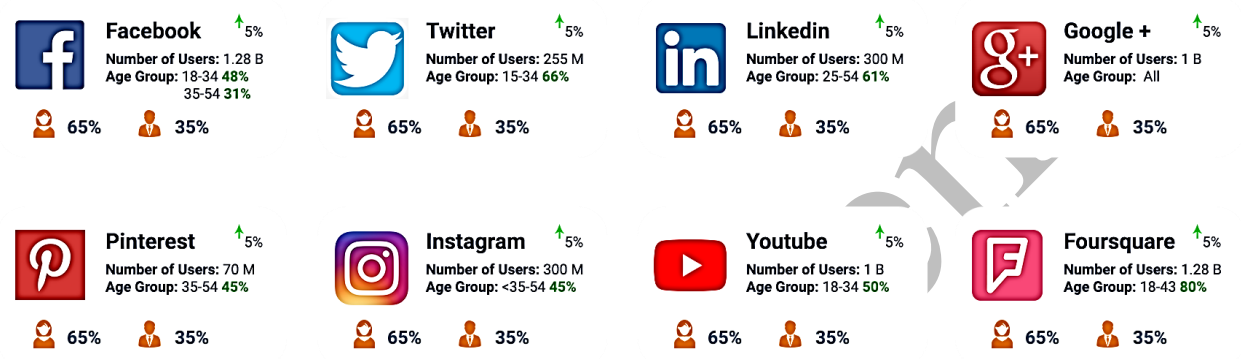
The graphical representation of data and information in the form of tables, charts, graphs & dashboards.

- Financial Reporting
- Operational Reports
- Historical Reports

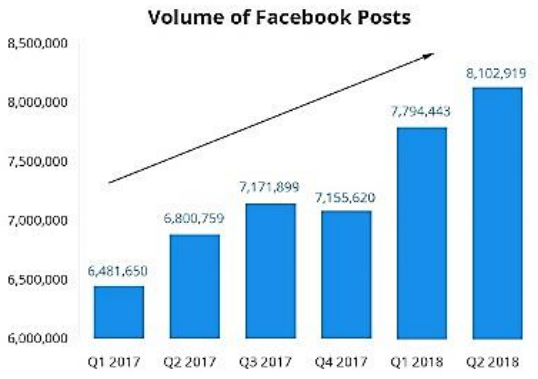
- Forecasting & Planning

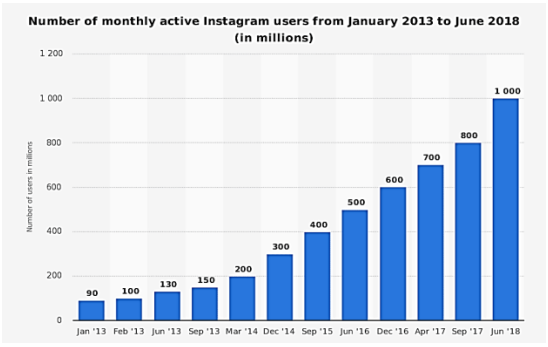
Module# 2: Data Science Hype

5 Billion+ Social Network Users

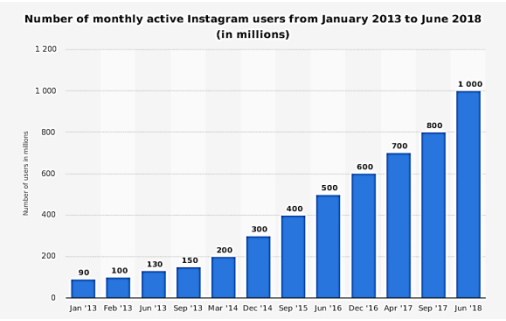


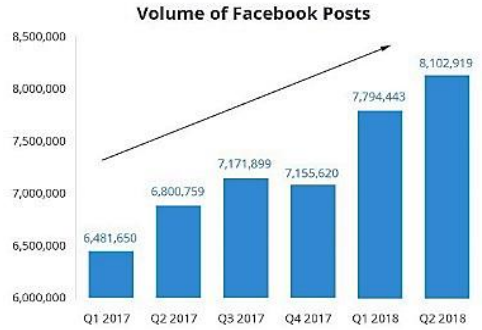
Analyze Social Network Trends





CS442- Introduction of Data Science

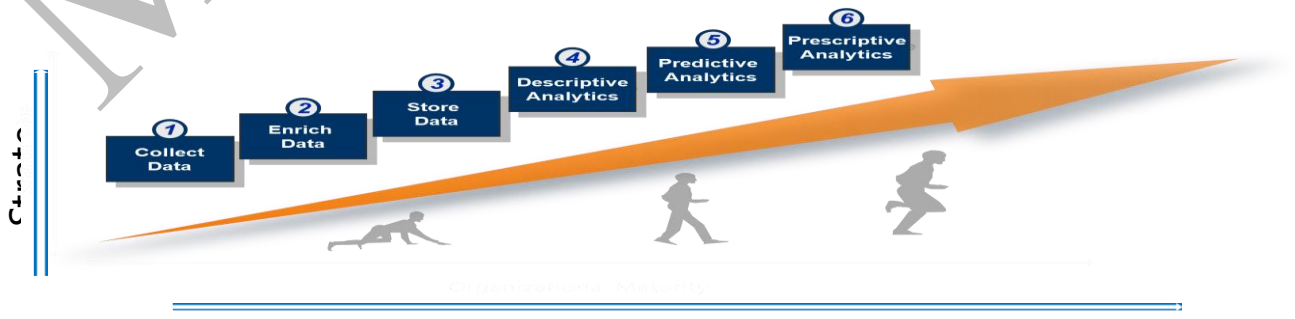




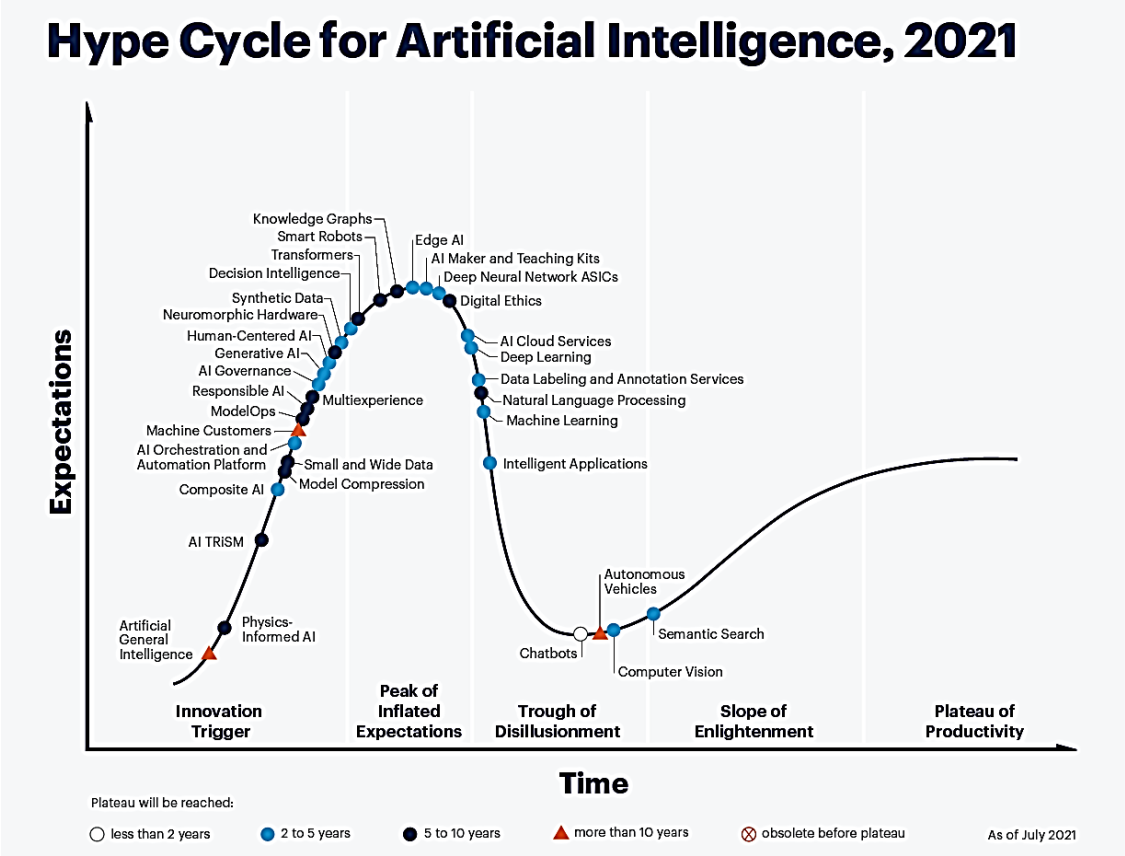
Examples of Social Network & e-Commerce Data

- Number of followers
- Follower Trends
- Number of likes
- Number of clicks
- Number of subscribers
- Number of customers
- Number of visitors

Organizational Data Strategy



CS442- Introduction of Data Science

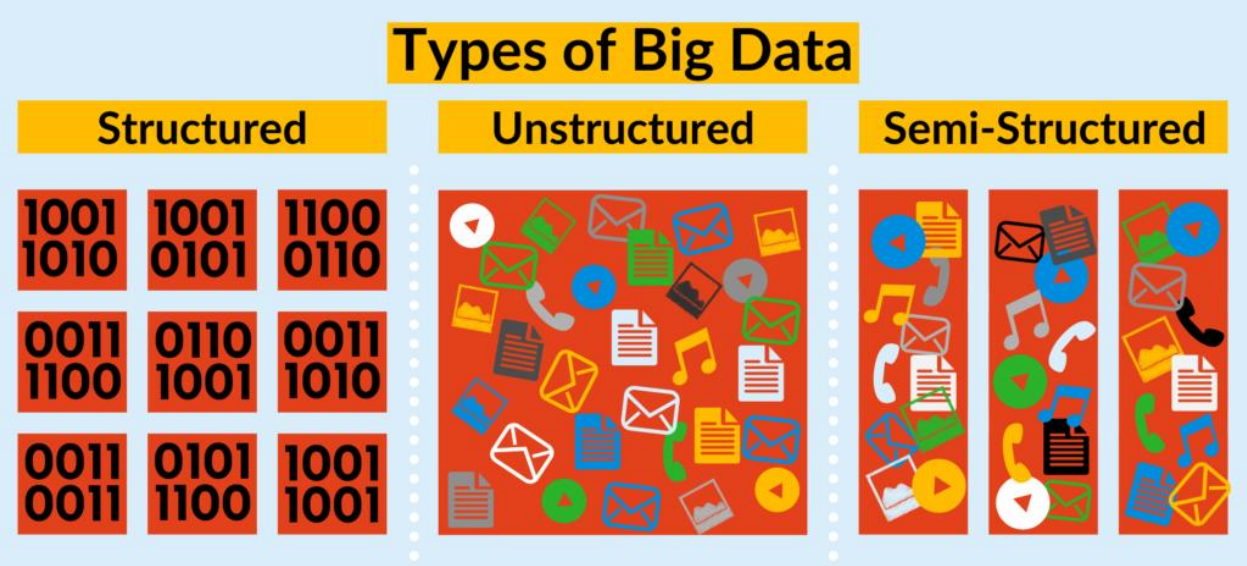


Module# 3: Big Data

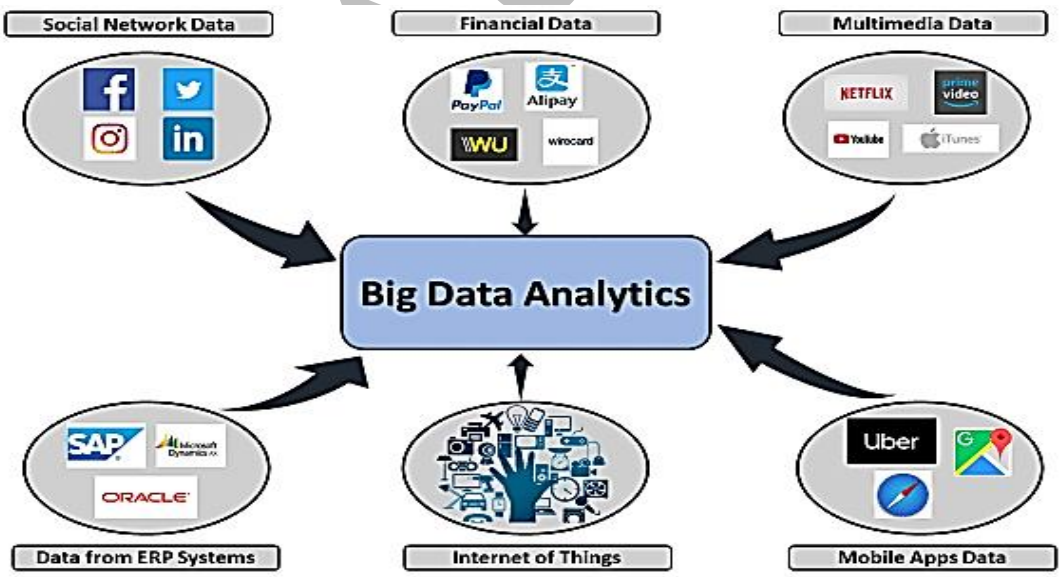
5 Vs of Big Data

- Velocity
- Volume
- Variety
- Veracity
- Value

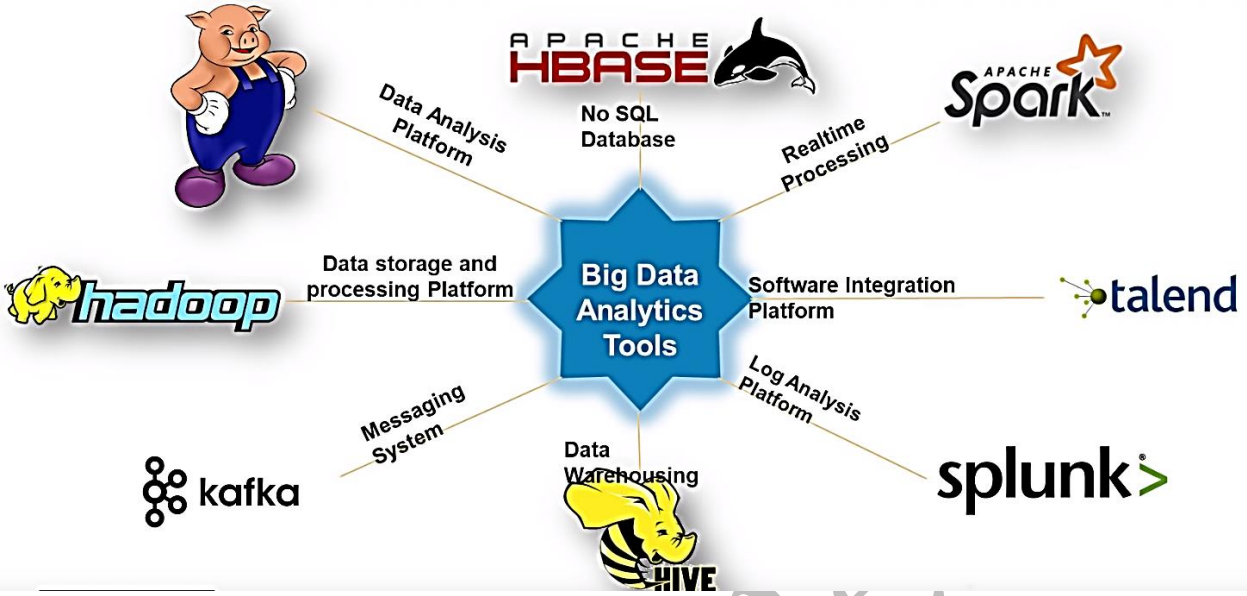
CS442- Introduction of Data Science



Sources of Big Data

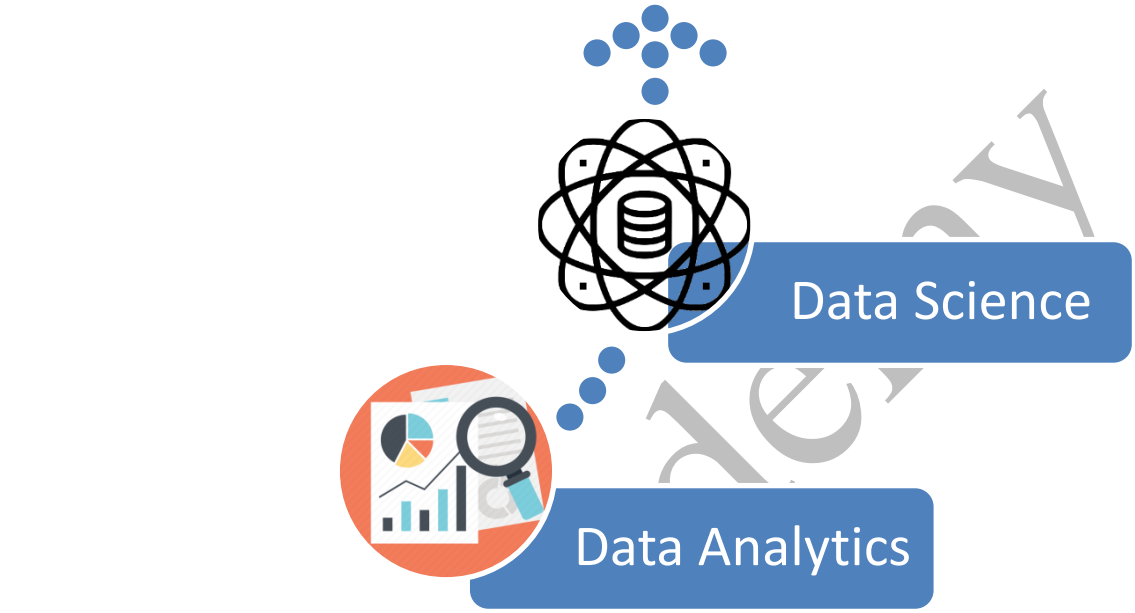


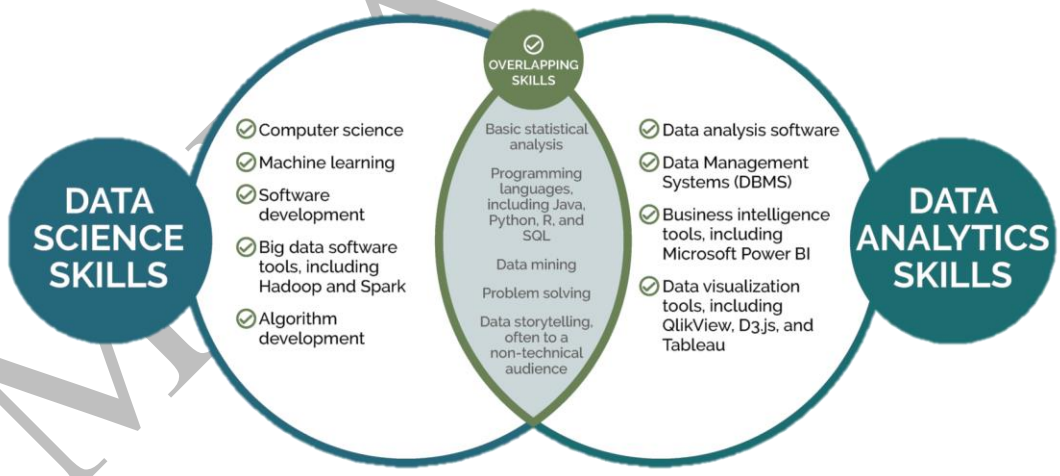
Big Data Technologies



Big Data Ecosystem

Module# 4: Data Science vs Data Analytics





Role of Data Scientist

Designing and development of new processes for data modeling and production using prototypes, algorithms, predictive models and custom analysis.

Data Science Use Case: Healthcare Diagnostics

- 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

Role of Data Analyst

- Examine large data sets to identify trends, develop charts, and create visual presentations to help businesses make more strategic decisions.

Data Analysis Use Case:

Electronics Health Records

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

Module# 5

What is Data Science: Datafication

Datafication Definition:

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

- Real-time tracking
- Data Monitoring
- Predictive Analysis and Optimization

CS442- Introduction of Data Science

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 a multitude of industries:

- **Google Maps**
- **Location Based Services**
- **Recommendation Engine**

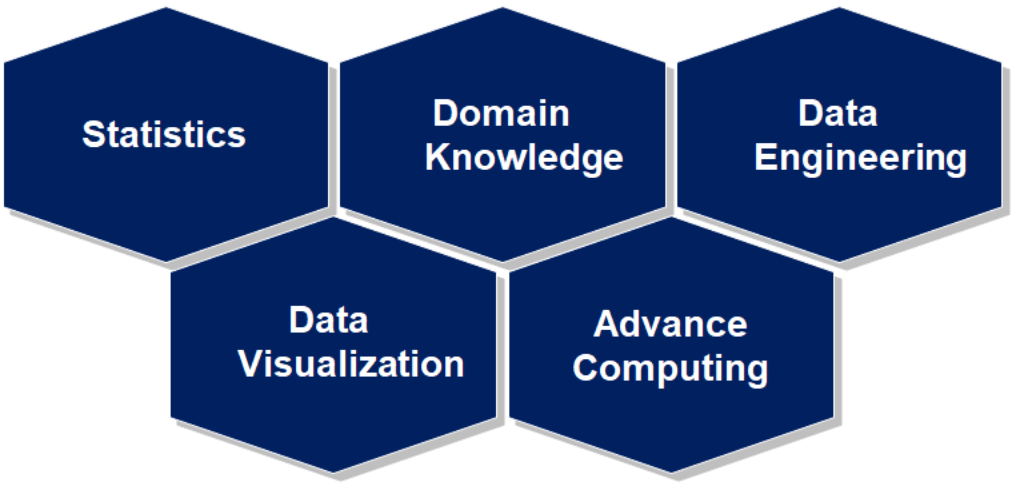
Module# 6: Current Landscape of Perspective

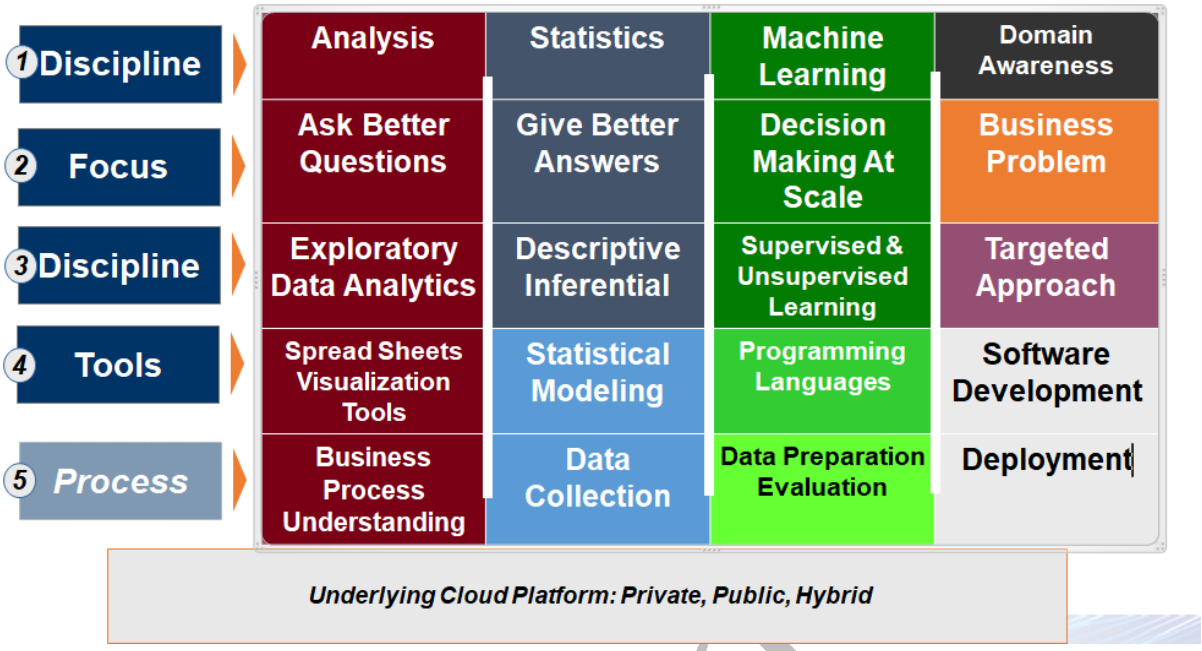
Data Growth Challenges:

- Data around us is growing at an exponential rate.
- Till 2025, global data creation is projected to grow to more than 180 zettabytes
- Just two percent of the data produced and consumed in 2020 was saved and retained into 2021
- Installed storage capacity is growing at CAGR of 19.2 till 2025

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.





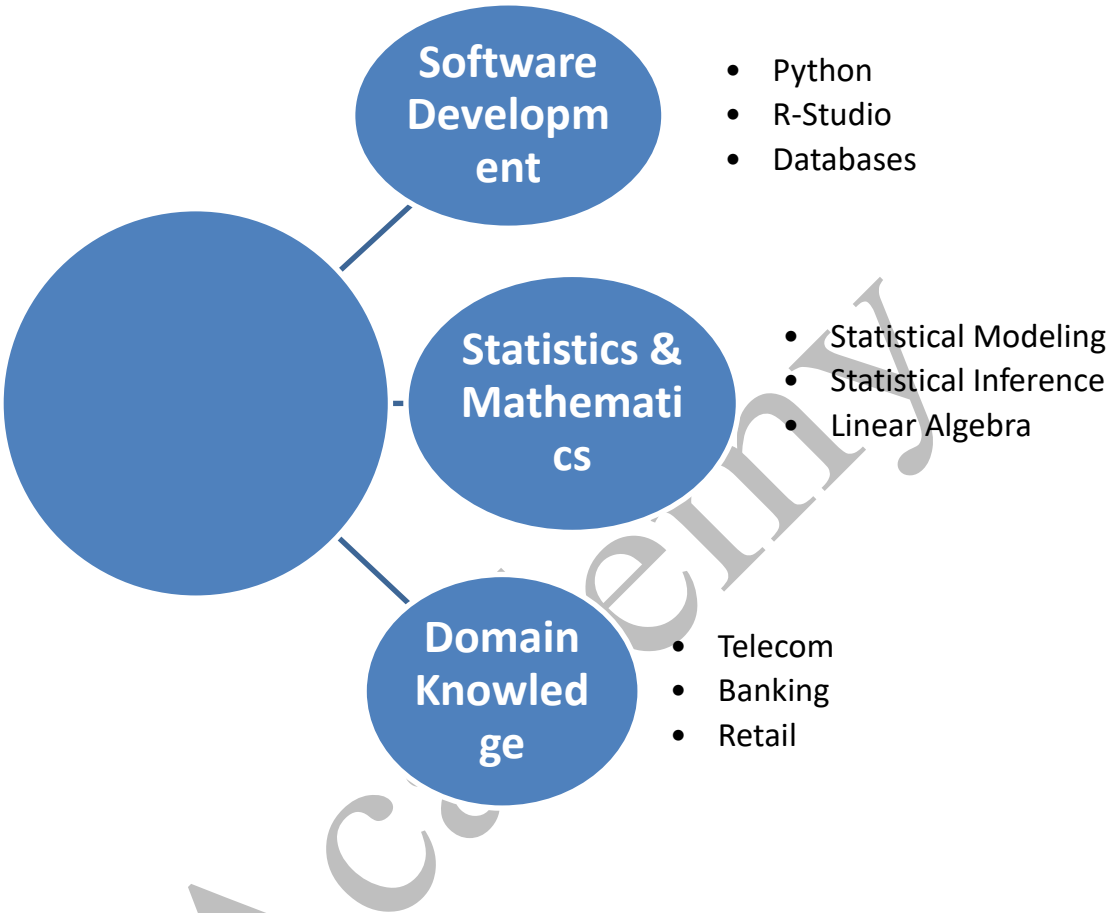
Module# 7: Skill Set Needed

- Data Science is about combining the extraction of knowledge from data to answer a particular problem allowing businesses and stakeholders to make intelligent data driven decisions.
- Complete technology stack together with soft skills brings more significant challenges to address.

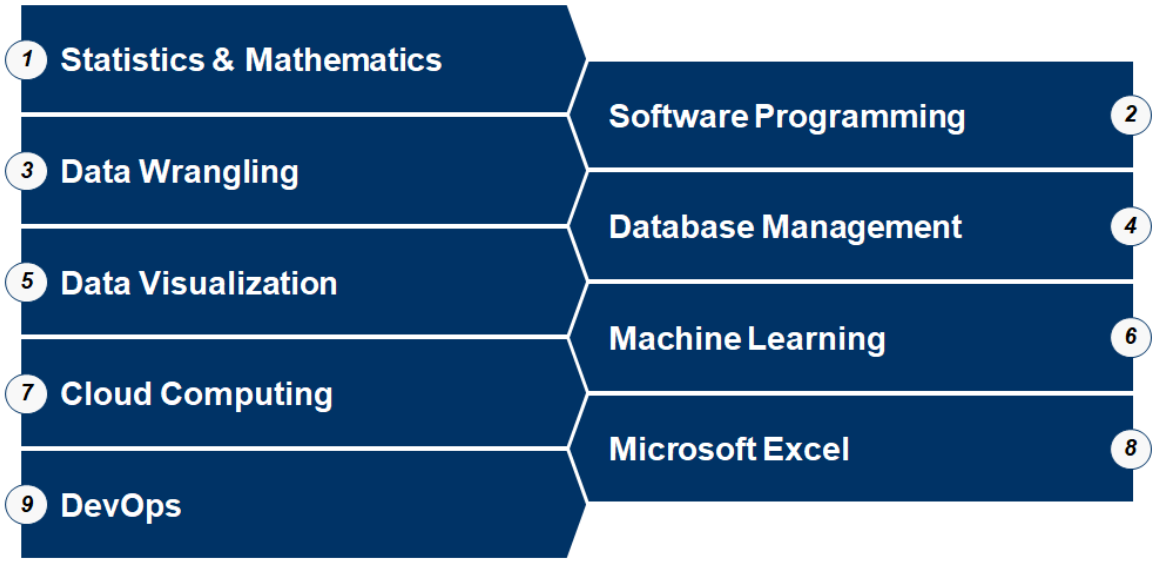
Data Scientist

- Designing and development of new processes for data modeling and production using prototypes, algorithms, predictive models and custom analysis.

CS442- Introduction of Data Science

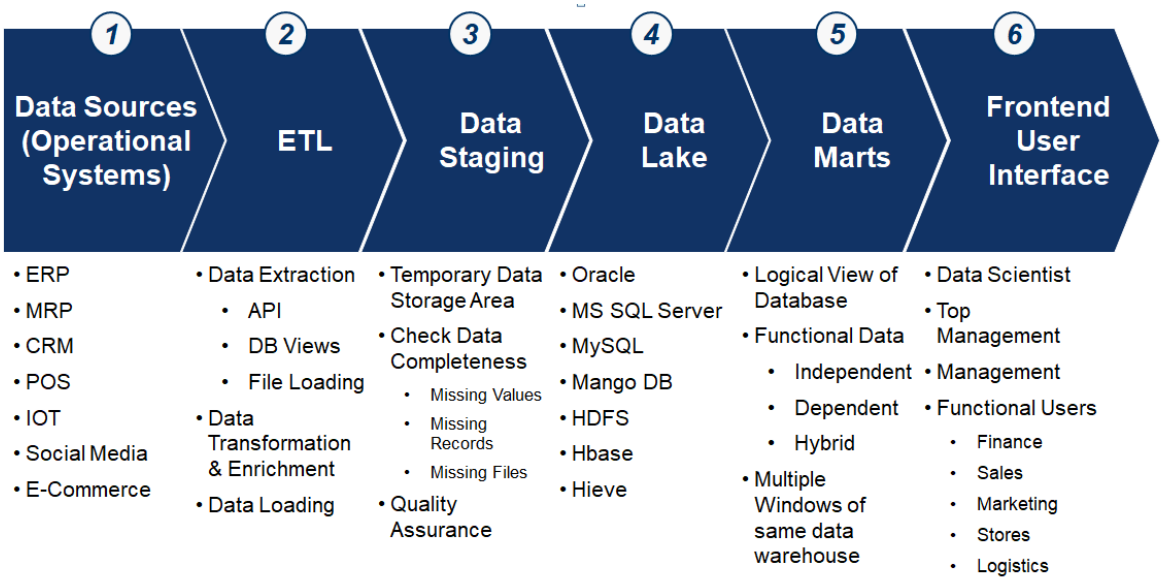


Organization



CS442- Introduction of Data Science

Module #8: Software Development



Object Oriented Programming:

- Object-oriented programming (OOP) 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 analytics codes into production following are few examples:

- Programming Languages
- Integration Services (Windows, XML, JSON)
- Big data – Hadoop, Mango DB
- DevOps – Jenkins, Dockers

.

CS442- Introduction of Data Science

Programming Languages:

Data science 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:

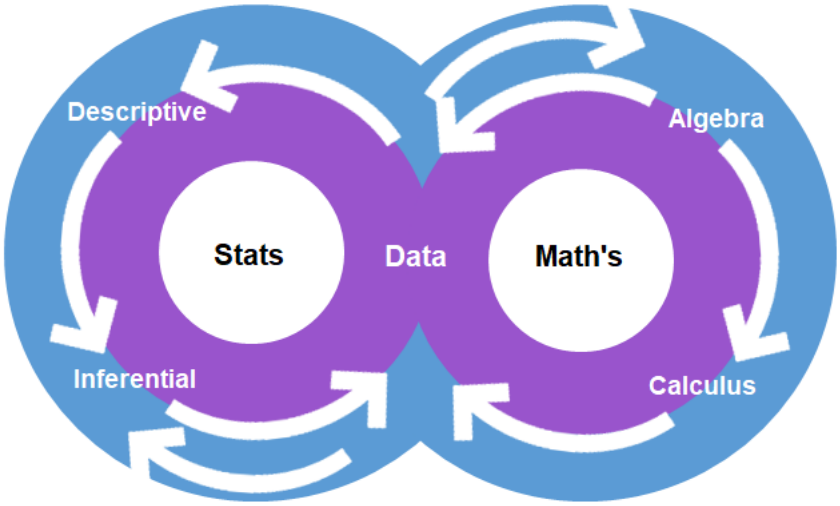
Data scientist need to analyze massive data sets and often require SQL and NoSQL tools & techniques.

- SQL Query Language
- Hieve
- HBASE
- SPARK

Module# 9: Statistics & Mathematics

Stats & Maths Essentials in Data Science:

Data Science revolves around these two fields and draws their methods and models to operate on the data.



Statistical Methods in Data Science:

- Descriptive Statistics works on facts about past data
- Inferential Statistics works on samples collected from the population to draw conclusions about the population

Mathematical Methods in Data Science:

Two main components of mathematics contribute to Data Science:

- Linear Algebra is widely used in image recognition, text analysis and also dimensionality reduction
- Calculus is used in optimization techniques for machine learning

CS442- Introduction of Data Science

Module# 10: Domain Knowledge

Why Its Important?

- Precise and accurate problem definition is critical for the overall success of a data analysis project. Domain knowledge helps to achieve precision and accuracy in derived results.

Subject Matter Expertise:

SMEs are considered the king! Especially with extensive experience working in a specific field it could potentially be the most valuable skill.

- They know source of problem, the business is trying to resolve
- The exactly know how the specific data is produced & collected

Best Domains for Data Science:

- Banking
- Insurance
- Healthcare
- Telecom
- Corporate Services
- Media & Communication
- Social Media & E-Commerce
- Software & IT Services

CS442- Introduction of Data Science

Module# 11: Introduction To Statistics

Mathematical Discipline

Including methods of:

- Data Collection
- Data Organization
- Data Analysis
- Data Interpretation
- Data Presentation

Examples of Statistical Data

- Height, Weight, Blood Pressure, Current, Voltage
- Number of Cars, Number of Children, Distance, Speed, Time, Pressure

Data Collection

- Surveys
- Census
- Sales
- Bank Transactions
- Online Shopping

CS442- Introduction of Data Science

Data Organization

Data Identification, Gathering & Structuring for Analysis:

- Tumor Registry Perform
- Motor Vehicle Registration
- Voters Registration List
- List of Tax Payers

Data Analysis

- Number of patients with a particular cancer type
- Number of Cars of a particular brand or capacity in a city
- Number of Registered Female Voters in a constituency
- List of Top 20% Tax Payers in a particular tax zone

Data Interpretation

- Demographics of cancer patients
- Relations between Tax Paid & Number of Cars owned
- Female Voter's Turnout in a particular districts
- Urban & Rural Divide between Tax Payers

Data Presentation

- Reports
- Charts

- Dashboards

In general statistical analysis falls into two categories:

- Descriptive Statistics
- Inferential Statistics

Module# 12: Descriptive Statistics

Descriptive Statistics

- Precisely Describe Collected Data
- Report Data Characteristics
    - Measures of Central Tendency
    - Data Spread
    - Relative Standing

CS442- Introduction of Data Science

Characteristics of Descriptive Analysis

- Describe summarize data
- Present data in meaningful way
- Describes already known data
- Data Tables, Charts And Dashboards are used
- Measures of central tendency, spread of data are the tools of descriptive statistics

Descriptive Data Example

- Math score for GCSE Students
- Collect Data For Last Three Years
- One Can Quickly Compare the results:
    - Number of Students with A Grade
    - Number of Students with B Grade
    - Number of Students with C Grade
    - Number of Students Failed

Data Certainties

- Precisely Describe Collected Data
- Report Data Characteristics
    - Measures of Central Tendency
    - Data Spread
    - Relative Standing

CS442- Introduction of Data Science

Measures of Central Tendency

- Mean
- Median
- Mode

Data Spread

- Variance
- Standard Deviation
- Correlation
- Covariance
- Range

Relative Standing

- Percentile
- Quartile
- Inter-Quartile Range

CS442- Introduction of Data Science

Module# 13: Inferential Statistics

Draw Conclusion About Population Based On Sample Data

- Generalize From Sample to Population
- Hypothesis Testing
- Making Prediction

Inferential Analysis

- Analysis of random sample of data taken to describe and make inference about the population
- Inferential statistics compares, test and predict the data
- Make conclusion about the population that is beyond the data available
- Probability scores are used to represent data

Inferential Analysis Tools **(Imp for MCQ + short)**

- The basic tools of inferential statistics are:
  - Hypothesis test
  - Analysis of variance(ANOVA)

Inferential Data Example

- Suppose you want to know the average height of all the men in a city with a population of so many million residents. It isn't very practical to try and get the height of each man

Population & Sample **(Imp for short question)**

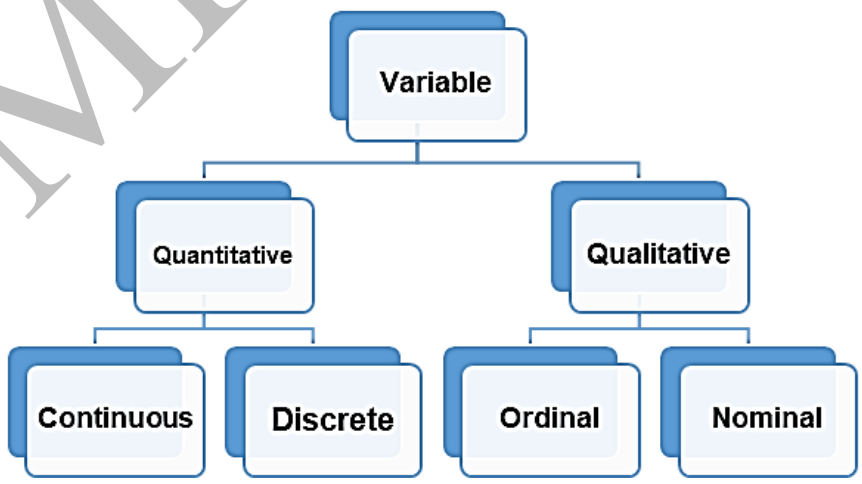
- Population: Total Number of Voters in a Constituency
- Sample: Number of voters selected for opinion pool
- Sample is a subset of the population **(imp for Mcq)**

Measure	Population Symbols	Sample Symbols
Size	N	n
Mean	□	x
Standard Deviation	δ	s
Percentage	π	p

Module-14: Statistical Variables

Statistical Variables

- A variable is a data item: characteristics, number, or quantity that can be measured or counted.
- It is called a variable because the value may vary between data units in a population, and may change in value over time.



CS442- Introduction of Data Science

Quantitative Variables

A variable is quantitative if its values or categories consists of numbers and if differences between its categories can be expressed numerically. Examples:

- Height of an individuals
- Age of an individuals
- Population size of city
- Number of square feet in a house
- Number of students in a class

Quantitative Variables

Discrete Variables:

- A variable whose attribute are separate from one another. It is A finite number of variable. Discrete variables is also known as qualitative variables. Discrete variable can be measured as Nominal and Ordinal scale.

Continuous Variables:

- A variable can be on any value between two specified values. It is an infinite number of values. Continuous variables is also known as quantitative variables. Continuous variables are measured as interval and ratio scale.

Qualitative Variables

Qualitative variables have discrete categories, usually designated by words, labels and nonnumeric differences between categories.

- Gender
- Breed of dog
- Level of education
- Marital status
- Eye color

Ordinal Variables:

- Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories are not known.

Ordinal Variable Examples:

- Customer Satisfaction Level
    - Happy
    - Satisfied
    - Some What Satisfied
    - Not Satisfied
    - Unhappy

Ordinal Variable Examples:

- Frequency of Going To Gym
    - Never
    - Rarely
    - Sometimes
    - Often
    - Regular

Nominal Variables:

- A nominal scale describes a variable with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.

Nominal Variable Examples:

- Gender

- Race
- Eye Color
- Language
- Marital Status

Data Unit	Numeric Variable	Quantitative Data	Categorical Variable	Qualitative Data
A Person	How many children do you have?	4 Children	In which country your children were born?	Pakistan
A House	How many square meters is the house ?	200 Square meters	In which city the house is located?	lahore
A Business	How many workers are currently employed?	264 Employees	What is the industry of business?	Retail
A Farm	How much milk cows are located on the farm?	36 cows	What is the main activity of the farm?	Dairy
An occupation	How much do you earn?	60,000 RS	What is your occupation?	Business man

Module-15: Fundamentals of Descriptive Statistical

To organize, summarize simplify, describe and present data in a meaningful manner. Following are the three main description classifications:

- Measures of Central Tendency
- Spread of Data
- Relative Standing of Data
-

Measure of Central Tendency:

- Mean
- Median
- Mode

eng50, 60math, urdu70

Sum = 180
Total values = 3

Mean = 180 ÷ 3 = 60 average marks = 60

CS442- Introduction of Data Science

Mean:

- In mathematics and statistics, the arithmetic mean, or simply the mean or the average, is the sum of a collection of numbers divided by the count of numbers in the collection. The collection is often a set of results of an experiment or an observational study, or frequently a set of results from a survey

Median:

10, 20, 30, 40, 50

- In statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution. For a data set, it may be thought of as "the middle" value.

Mode:

10, 20,20, 30, 40, 50

- The mode is the value that appears most often in a set of data values. If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled

Spread of Data:

- Standard Deviation
- Variance
- Covariance
- Correlation
- Range

Standard Deviation:

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Variance:

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its population mean or sample mean. Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value.

CS442- Introduction of Data Science

Covariance:

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariance is positive.

Correlation:

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it actually refers to the degree to which a pair of variables are linearly related.

Range:

In statistics, the range of a set of data is the difference between the largest and smallest values. Difference here is specific, the range of a set of data is the result of subtracting the sample maximum and minimum. However, in descriptive statistics, this concept of range has a more complex meaning.

Relative Standing of Data:

Measures of relative standing, in the statistical sense, can be defined as measures that can be used to compare values from different data sets, or to compare values within the same data set.

- Percentile
- Quartile
- Inter-Quartile Range

Percentile:

- In statistics, a k-th percentile, is a score below which a given percentage k of scores in its frequency distribution falls or a score at or below which a given percentage falls. For example, the 50th percentile is the score below which or at or below which 50% of the scores in the distribution may be found

Quartile:

- A quartile is a statistical term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations.

Inter-Quartile Range:

- In descriptive statistics, the interquartile range is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the midspread, middle 50%, or H-spread. It is defined as the difference between the 75th and 25th percentiles of the data.

Module- 16: Fundamentals of Inferential Statistics

Fundaments of Inferential Statistics are to:

Use sample data to make inferences and generalization about the whole population. Following are main concepts:

- Population & Samples
- Probability Theory & Distributions
- Estimation
- Central Limit Theorem
- Hypothesis Testing

Population & Samples (important topic)

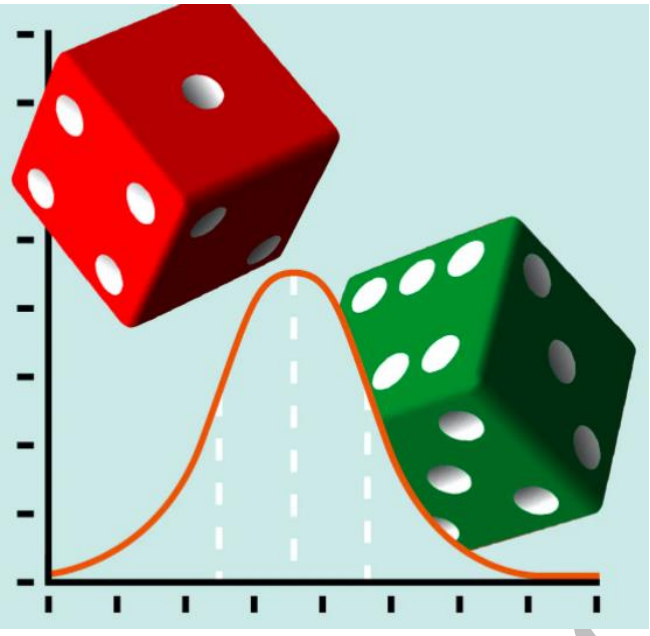
A variable is quantitative if its values or categories consists of numbers and if differences between its categories can be expressed numerically. Examples:

- Height of individuals
- Age of individuals
- Population size of city
- Number of voters
- Number of students in a class

CS442- Introduction of Data Science

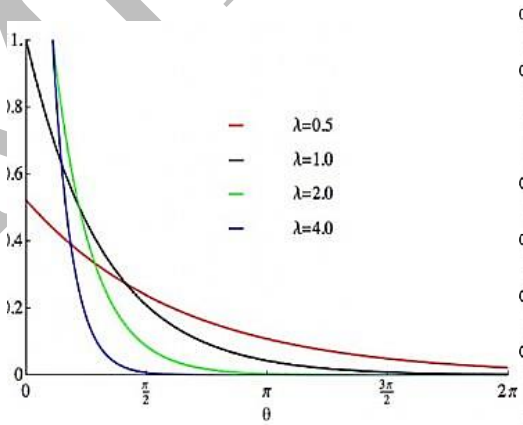
Probability (MCQ)

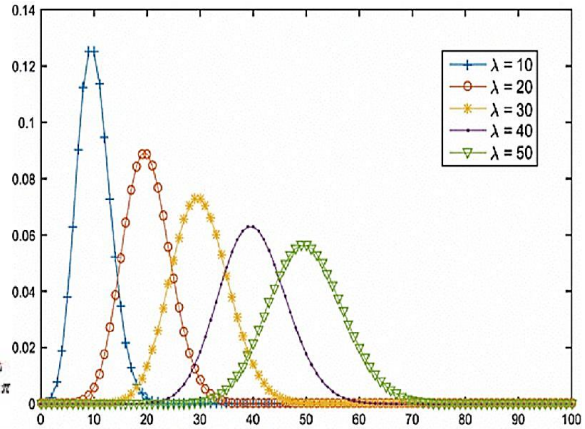
Probability theory is the branch of mathematics concerned with probability. Although there are several different probability interpretations, probability theory treats the concept in a rigorous mathematical manner by expressing it through a set of axioms.



Probability Distribution (MCQ)

Probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events.





CS442- Introduction of Data Science

Estimation

Statistical estimation is a function of the sample for calculating an estimate of a given quantity based on observed data: thus the rule, the quantity of interest and its result are distinguished. For example, the sample mean is a commonly used estimator of the population mean.

Central Limit Theorem

The central limit theorem establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed

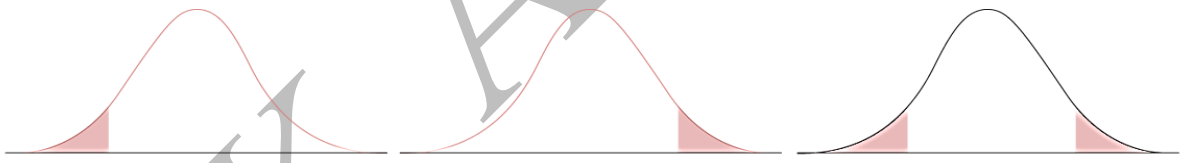
Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to determine a possible conclusion from two different, and likely conflicting, hypotheses. In a statistical hypothesis test, a null hypothesis and an alternative hypothesis is proposed for the probability distribution of the data.

$H_o : \mu = claim$
$H_1 : \mu < claim$

$H_o : p = claim$
$H_1 : p > claim$

$H_o : \beta = claim$
$H_1 : \beta \neq claim$



Module-17: Statistical Inference:

Statistical Modeling

Statistical modeling is the process of applying statistical analysis to a data set, embodies a set of statistical assumptions concerning the generation of sample data. A statistical model represents the data under observation.

Statistical Model Types:

There are three main types of statistical models:

- Parametric Models

- Non-Parametric Models
- Semi-Parametric Models

Parametric Models

- A parametric model or parametric family is a finite-dimensional model. It is a family of probability distributions that has a finite number of parameters.

Parametric Models Example:

- Normal Distribution
- Poisson Distribution
- Exponential Distribution
- Linear Regression
- ANOVA

Non-Parametric Models

- The Non-parametric Models are statistical models that do not often conform to a normal distribution, as they rely upon continuous data, rather than discrete values. Non-parametric statistics often deal with ordinal and nominal numbers.

Non-Parametric Models Example:

- Chi-Square
- Sign Test
- Friedman Test
- Wilcoxon Signed Rank Test
- Mood's Median Test

Semi-Parametric Models

- A semiparametric model is a regression model with both parametric and nonparametric components used for estimation of finite-dimensional parameters in a model.

Semi-Parametric Models Example:

- Normal Distribution
- Linear Regression
- Logistic Regression
- Cox Proportional Hazards
- Gaussian Mixture Models

Issues with Statistical Models:

- Different experiments yield different parameter estimates
- A parameter estimate has a sampling distribution
- Better parameter estimates will have sampling distributions that are closer to the true parameter
- Given a parameter estimate, how well does the distribution fit the data?

Module-18: Statistical Inference: Fitting a Statistical Modeling

Model fitting is a procedure that takes three steps:

- Select a function that takes in a set of parameters and returns a predicted data set
- Select an 'error function' that provides a number
- To evaluate “Goodness of fit” compare observed data to expected data

Poisson Distribution Model:

- Data consists of counts of occurrences
- Events are independent
- Mean number of occurrences stays constant during data collection
- Occurrences can be over time or over space (distance/area/volume)

Poisson Distribution Examples:

- Number of incoming phone calls in a telecom company call center
- Individuals diagnosed with a rare disease in a community
- Number of accidents on a particular route or section of the road
- Number of hits on a website

Transport Modeling:

- Journey time forecasts are required in many new dynamic traffic control and route guidance systems. This forecasting model is for urban networks for Normal and Incident conditions.

Four-Stage Model Structure

1. Trip generation
2. Trip distribution
3. Mode choice
4. Assignment

Transport Modeling

Study Area Zones

- Residential
- Commercial
- Industry
- Retail

Trip Generation

- Work
- Education
- Shopping
- Recreation
- Others

Transport Mode

- Own Car
- Bus
- Taxi
- Metro

Trip Distribution:

Linking trip generation to trip attractions:

Typically depicted using a "Gravity Model" function

$$T_{ij} = \acute{\alpha}\, O_i\, D_j\, f\,(C_{ij})$$

where T_{ij} *is the number of trips between zones I and j*

$\acute{\alpha}\, O_i\, D_j$ *are measures of size of origin zone i and destination zone j*

$f\,(C_{ij})$ *is a measure of the deterrence (cost) between zones i and j*

Transport Modeling Summary:

- Transport modelling allows the synthesis of travel behaviour and the prediction of future demands
- The four-stage approach is still the norm but with multiple variations in model design and structure
- In all cases, the model is still only a partial depiction of some elements of real life

Module# 19: Introduction To Cloud Computing

What is Cloud

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

- Data Storage
- Computing Power
- RAM
- Internet Connectivity

Without direct active management by the user.

What is Cloud

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

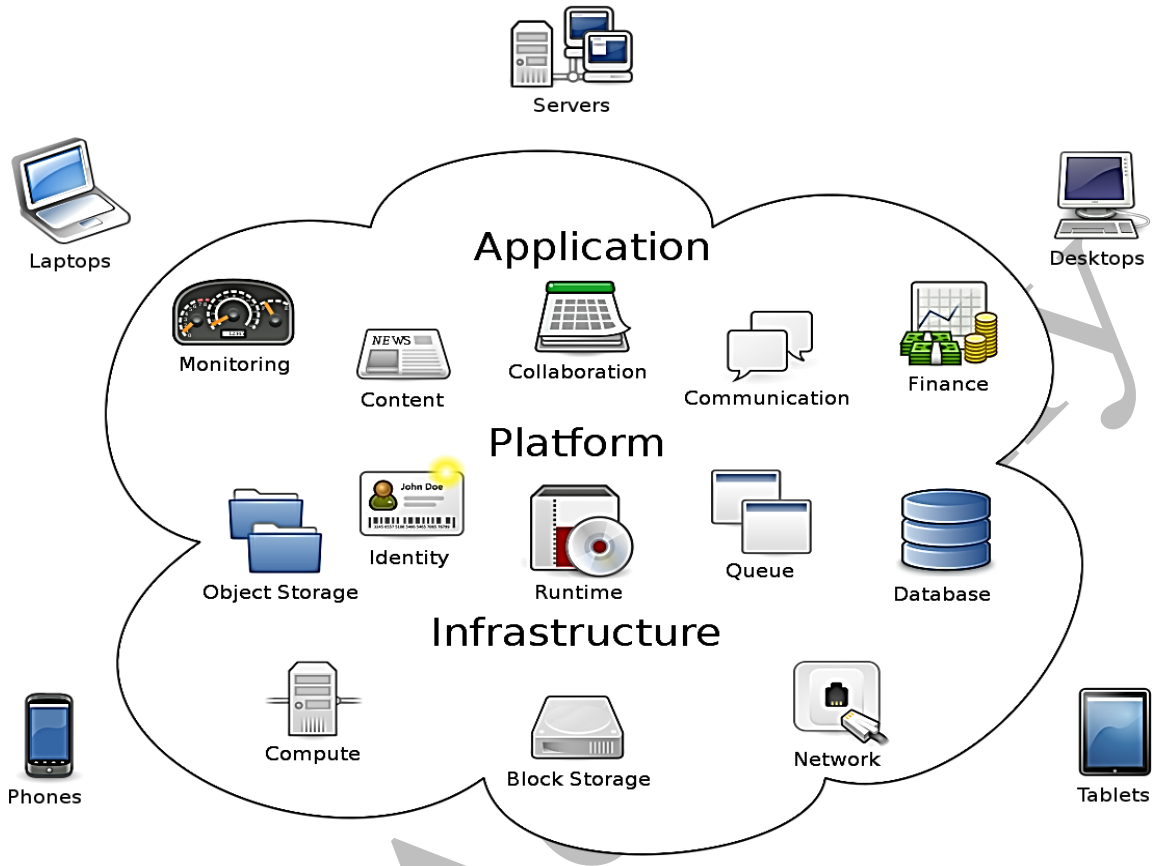
- Private Cloud
- Public Cloud
- Hybrid Cloud

Cloud Service Models:

Clouds operators often have functions distributed over mult:

- SaaS – Software As A Service
- DaaS – Data As A Service
- PaaS – Platform As A Service
- IaaS – Infrastructure As A Service

Cloud Infrastructure



Module# 20: Data Science @Cloud

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

Why Its Important:

Analysis and storage are the two important challenges for organizations large and small.

- Amount of Big Data generated has accelerated tremendously
- Storing data in cloud is:
    - Economical

- Secure

Cloud Computing Benefits For Data Scientist:

- Use platforms for free or at a minimal price
- Choice of available tools:
  - Open Source

  Hadoop, Python, MySQL
  - Commercial

Oracle, MS SQL, SAS, Tableau

Module# 22: Google Cloud Platform

What is Google Cloud:

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

- Google Search
- GMail
- Goggle Drive
- YouTube
- Google Translate

Google Cloud Benefits:

- Highly Productive & Innovative
- East Adoptability
- Remote Access
- Unmatched Security

- Flexibility & Control
- Allows to Store & Compute Data
- Manage SDLC
- API's Available

Google Cloud Platform

Module# 23: 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:

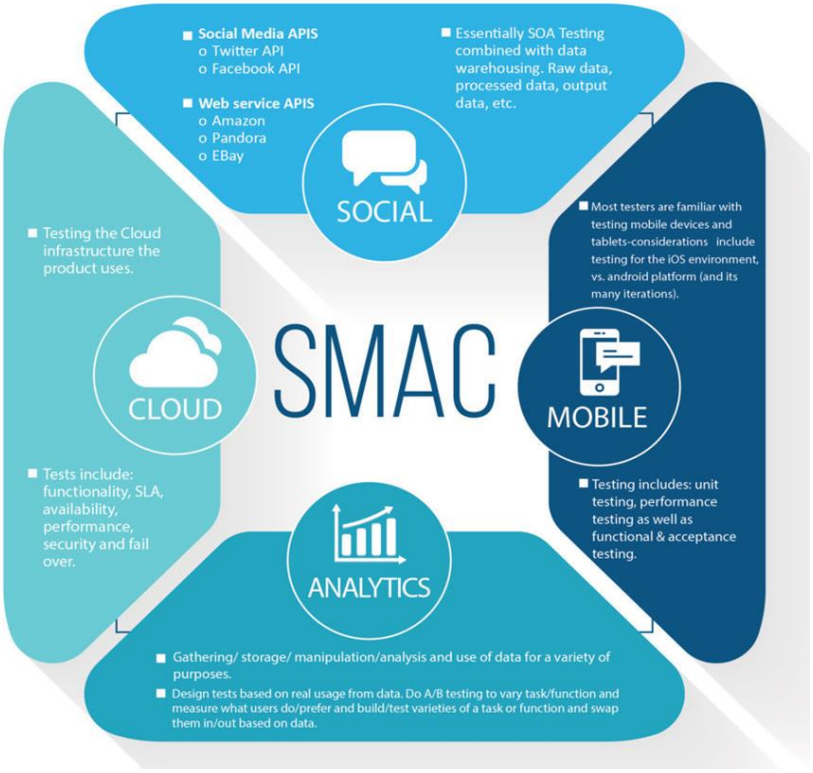
- Cost Optimization
- Unmatched Speed of Services
- Global Scale Digital Workplace
- Productivity & Performance
- Flexibility, Reliability & Security
- CICD Pipelines (DevOps)
- Disaster Recovery, Business Continuity
- Analytics
- Azure APIs & Workflows

Module# 24: Social Media Analytics on Cloud

Social, Mobile, Analytics & Cloud:

SMAC is an ecosystem enabling business transformation from e-business to digital and improve business operations and with minimal overhead and maximum reach:

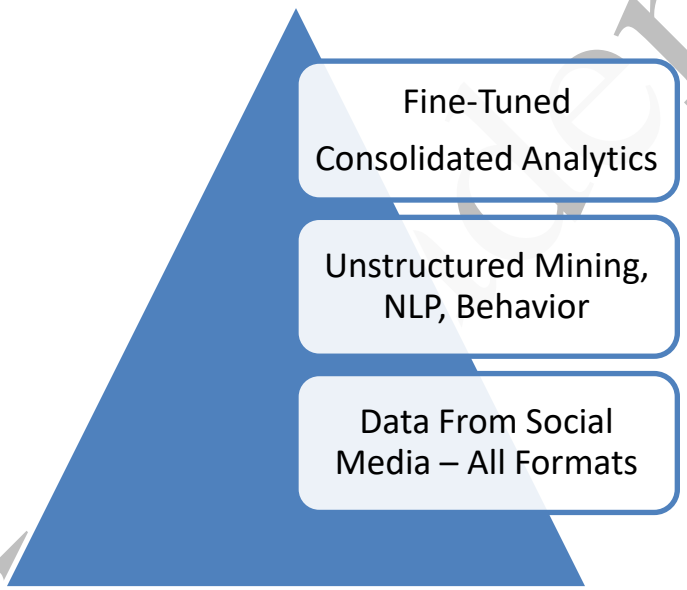
- Social Media
- Mobile Applications
- Data Analysis
- Cloud Computing



Social Media Analytics on Cloud Offers the following benefits:

- Integrated Data Storage

- Cost Effective
- Robust & Fast Computations
- Multi Tasking
- Statistical Modeling
- AI/ML Algorithms
- Actionable Insights



Module# 25: Introduction to Big Data

Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them.

What Is Big Data:

- Volume

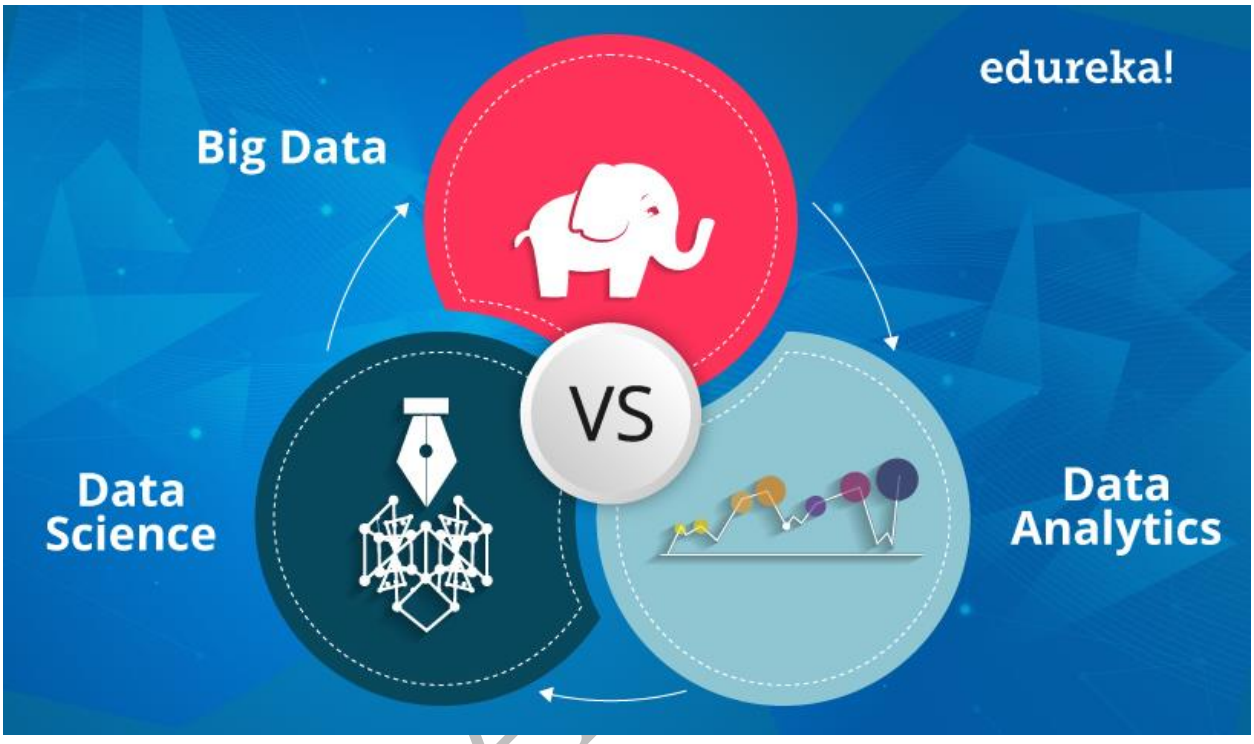
CS442- Introduction of Data Science

- Variety
- Velocity
- Veracity
- Value

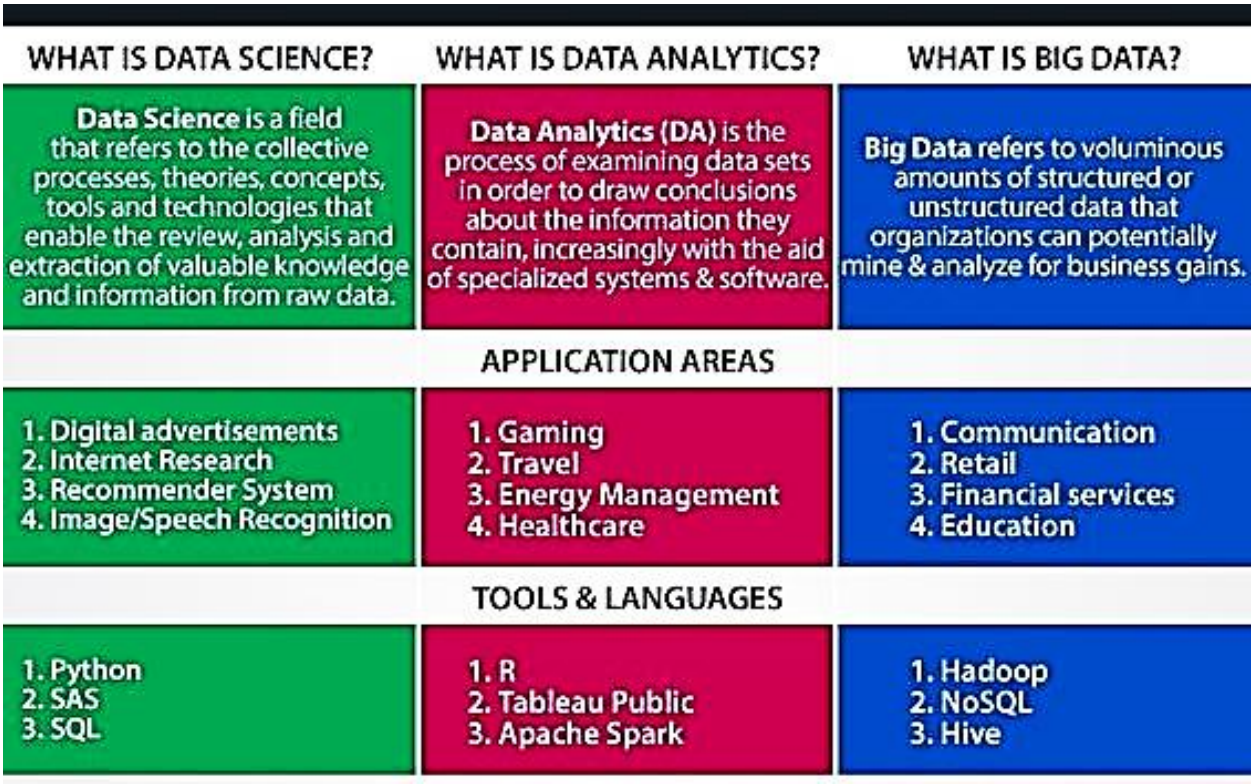
CS442- Introduction of Data Science

Module# 26: Data Analytics:

Big Data Vs Data Science Vs Data Analytics



CS442- Introduction of Data Science



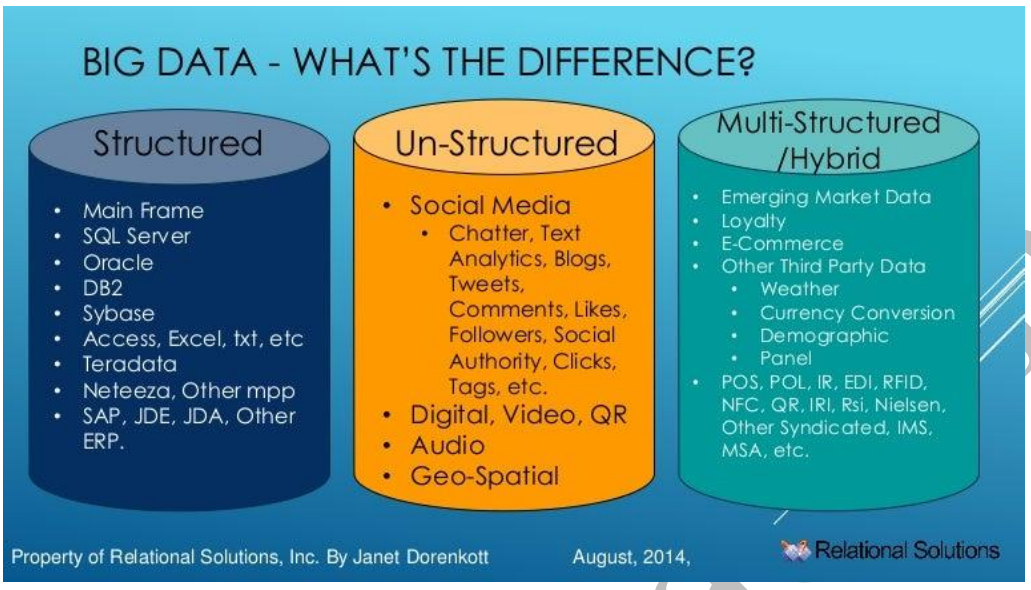
The World is generating Data at a higher rate, so the need of:

- Data Science
- Data Analytics

Is increasing to analyze and manage Big Data!

CS442- Introduction of Data Science

Module# 27: Data Types



Big Data Classifications:

- Structured Data
- Un Structured Data
- Semi Structured Data

Structured Data:

It adheres to a pre-defined data model and is therefore straightforward to analyze:

- Tabular
- Relationship between rows & columns
- Data Structure

Un-Structured Data:

Information that is not arranged according to a pre-set data model or schema, and cannot be stored in tables:

- Social Media

- Email, Documents
- Video
- Photos, Images
- IOT Logs, Censors

Semi-Structured Data:

The data that is not captured or formatted in conventional ways and does not follow the format of a tabular data model and does not have a fixed schema:

- XML Files
- TCP/IP Packets
- ZIP Files
- CSV, JSON

Module# 28: Handling Unstructured Data

How To Manage?

It is the process of collecting, storing, organizing, and analyzing data that doesn't have any predefined structure:

- Identify Data Sources
- Access & Organize Data
- Clean Data
- Analyze Data
- Visualize Data

Unstructured Data Challenges:

Including methods of:

- Data Format
- Data Quality
- Data Silos

- Data Growth

Unstructured Data Management Benefits:

Having a solid data management strategy helps to overcome big data management challenges and lead to benefits:

- Increased Productivity
- Accurate & Fast Decisions
- Better Compliance
- Improved Data Security
- Visualize Data

Understanding Big Data Technologies:

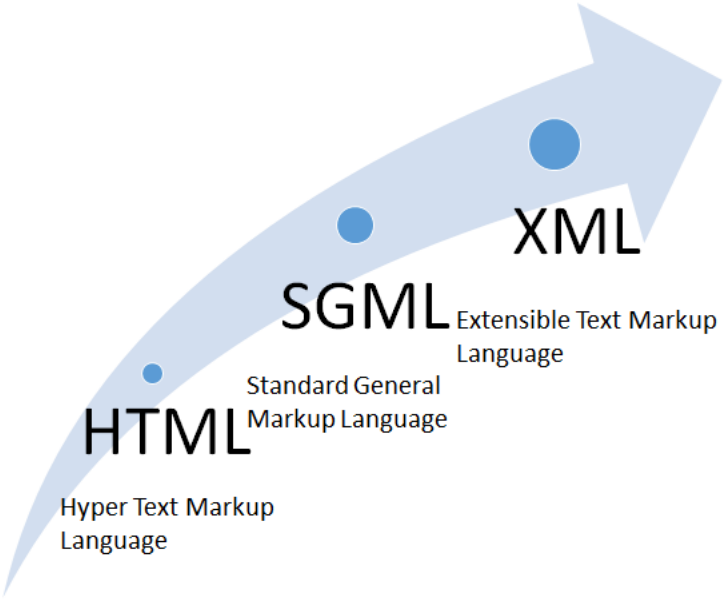
- Hadoop (HDFS)
- Machine Learning
- Natural Language Processing (NLP)
- Data Lake
- Monkey Learn
- RAPID Miner
- KNIME

Module# 29: Understanding XML

What is XML

XML Objects are universal data structure for big data and converts unstructured data to well-formed structure. It is widely used in software development and deployment context:

- Designed to describe Data
- Use XML Schema
- Store & Transport Data
- Publishing & Interchange



XML Applications

XML Objects are universal data structure for big data and converts unstructured data to well-formed structure. It is widely used in software development and deployment context:

- E-Commerce
- IOT
- ETL/ELT
- Edge Analytics

XML Databases

XML Objects are universal data structure for big data and converts unstructured data to well-formed structure. It is widely used in software development and deployment context:

- eXist DB
- BaseX
- Berkeley DB
- Sedna

XML (Extensible Markup Language)	HTML (Hypertext Markup Language)
It stores and transports data.	It displays data.
It uses user-defined tags.	It uses predefined tags.
It contains structural data.	It doesn't contain any structural data.
It can distinguish uppercase and lowercase letters (case sensitive).	It can't distinguish uppercase and lowercase letters (case insensitive).
It maintains spacing, tabs, newlines, and any other whitespace formatting.	It doesn't maintain whitespace.
It needs to have an end-tag.	It doesn't need an end-tag.
It needs structure or nesting.	It doesn't need structure.

Module# 30: Understanding JSON

JavaScript Object Notation:

It is a data-interchange format that is most commonly used to send data between web applications and a server:

- Its Scalable
- Its Lightweight
- Works with most of the modern programming languages
- Text based – Easy to Read & Write

JSON Benefits:

It is a superior format for WEB APIs and Web Development both at Front End & Back End. Is has interfaces with:

- JavaScript

- PHP
- Node JS
- Python
- C++

JSON Advantages:

It is generally used in the REST request and response API services, as JSON is uncomplicated and in a readable format, it is:

- Faster
- Support Schema
- Support Namespace
- Browser Compatibility
- Server Parsing
- Tool for Data Sharing

JSON	XML
Text based format (Not a Language)	Markup Language
Free to define anything	Has some rules
Smaller Size	Big in Size due to markups
JSON is similar to Java script Objects literals. Browser read faster.	Browser need parsers to handle XML. Slow processing.
No support on namespaces and comments	Both are supported.

Module# 31: Understanding Hadoop

Apache Hadoop

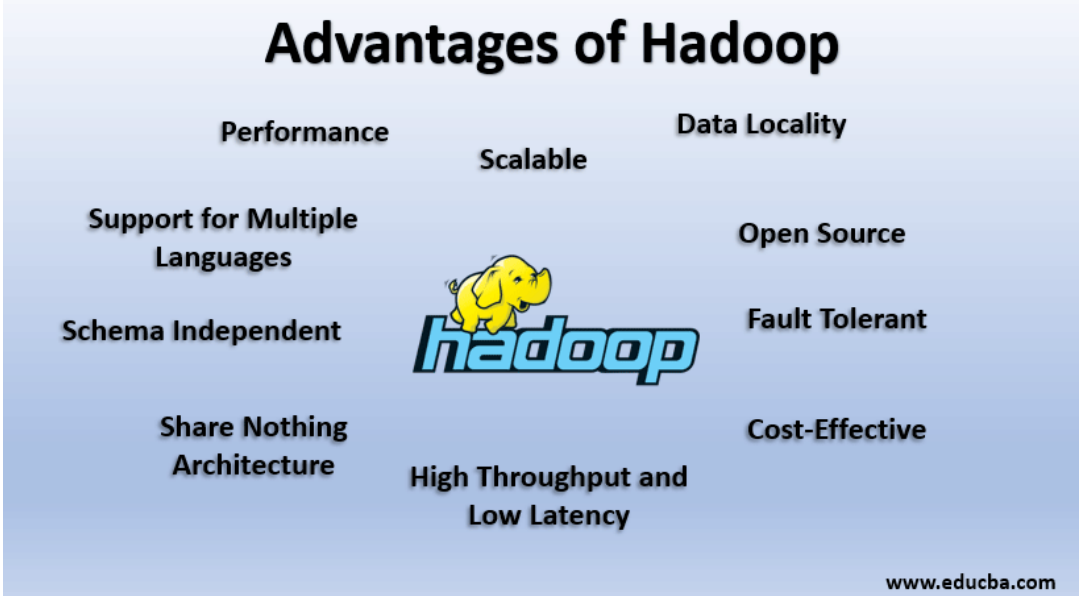
It is a popular Big Data Framework comprising of many Open Source utilities and two key modules:

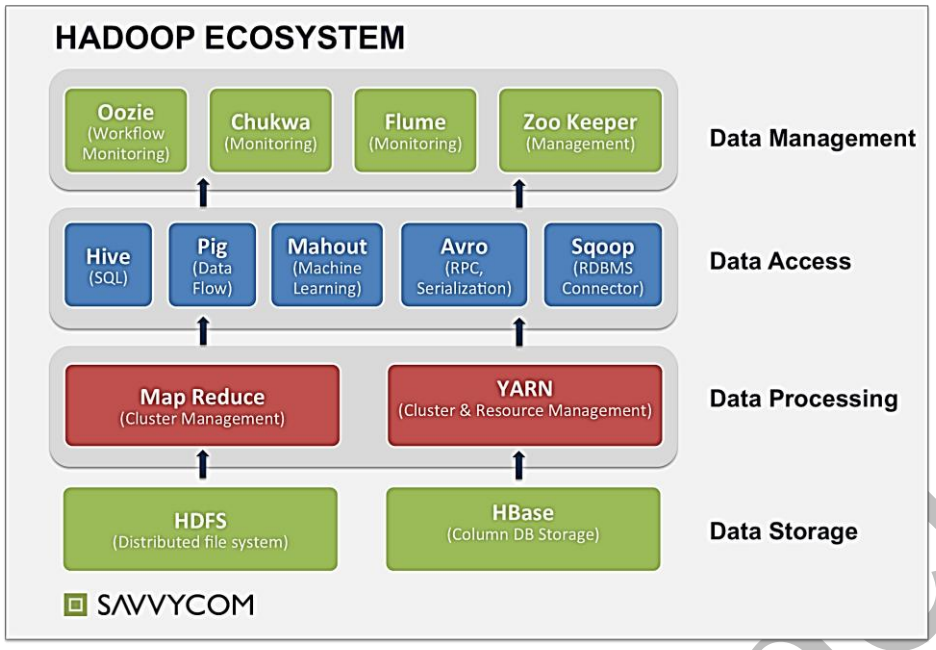
- HDFS (Hadoop File System)

Distributed File System for data storage

- MAP Reduce

Data Processing Framework





Module# 32: Understanding Spark

Apache Spark

It is an Open Source Unified Data Processing Engine:

- For Large Scale Data Processing
- Provide Interface For Programming Clusters (Hadoop)
- Parallel Data Processing
- Fault Tolerance
- Superior Processing Speed

Apache Spark Supports:

- Allow Programming in many languages (Python, Scala, Java)
- Built-in Operators (80+)
- SQL Queries
- Machine Learning
- Graph Data Processing
- In-Memory Intermediate Results

Spark Eco System:

- Spark Streaming

Process real time streaming data

- Spark SQL

Expose Spark Data Set over JDBC & APIS

- Spark Mlib

Machine Learning Library with built-in algorithms (Regression, Clustering, Filtering)

- Spark GraphX

API for Graphs & Graph Parallel Computations

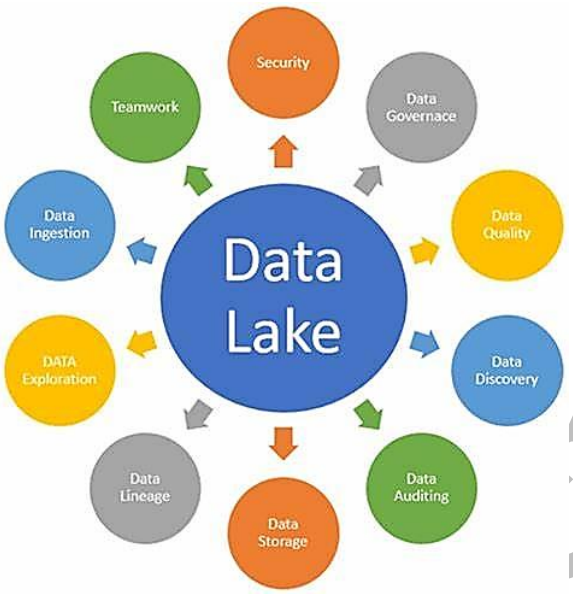
Module# 33: Data Lake

What is Data Lake?

It is a data repository including raw copies of source system data and transformed data used for tasks such as:

- Visualization
- Advanced Analytics

- Machine Learning



What is Data Lake?

A data lake can include all data formats:

- Relational Databases

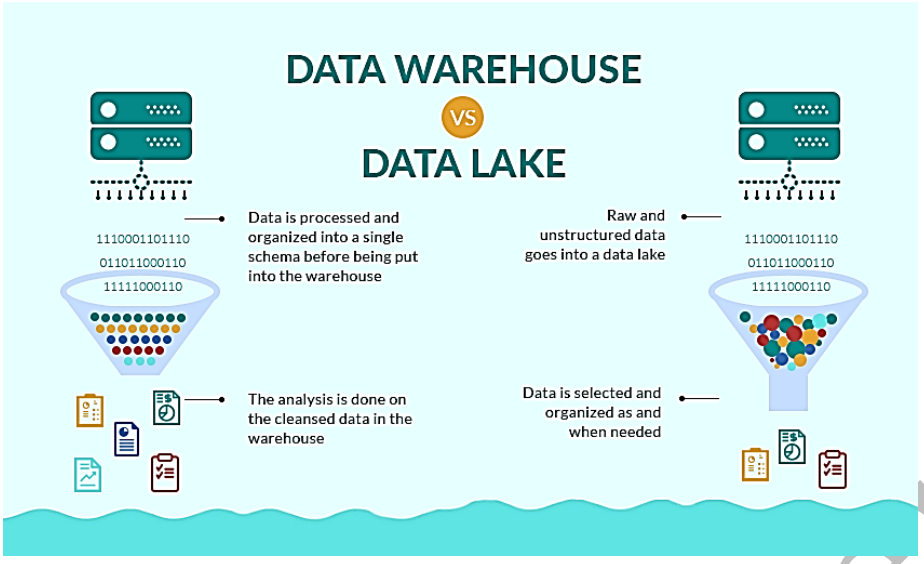
(Tabular: Rows & Columns)

- Semi-Structured Data

(CSV, logs, XML, JSON)

- Unstructured Data

(E-mail, Documents, Images, Audio, Video)



Data Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Module# 34: Healthcare Analytics

Health care analytics is the result of data collected from following areas within healthcare:

- Medical Claims and Cost

- Pharmaceutical and R&D Data
- Clinical Data (Collected from HER)
- Labs Data (Pathological & Radiology)
- Patient Behavior and Sentiment Data

Healthcare Data

- Relational Databases - EMR

(Patient Demography, Medical History)

- Semi-Structured Data - APIs

(Data from other Systems/Equipment)

- Unstructured Data - LABS

(Lab Reports, Images, Telemedicine)

Module# 35: Introduction to SQL

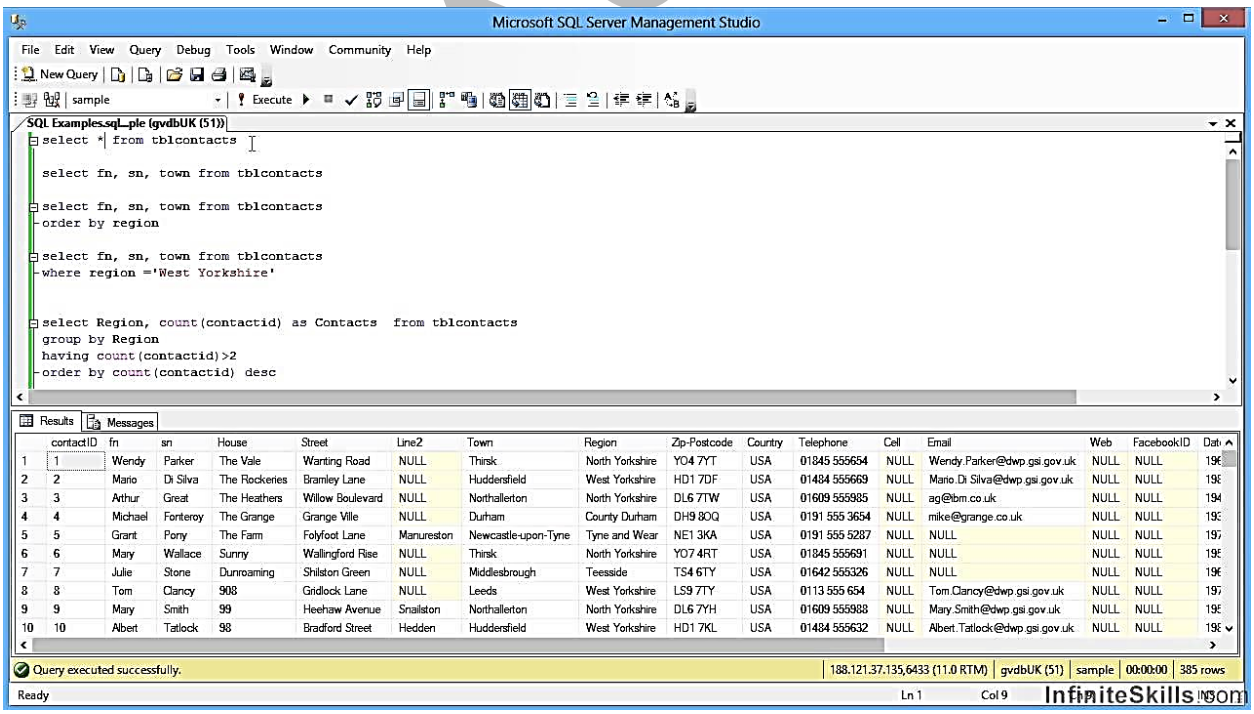
SQL Structured Query Language is one of the most demanding skill in Data Science used to perform:

- Create
- Read
- Update
- Delete

Four Basic Operations (CRUD) in Relational Databases.

Structured Data Programming Language:

- Structured Data is in the form of tables (Rows & Columns)
- Columns Describes Data Properties
- Rows Describe Real Data Entries
- Columns and Tables are related to each other
- SQL is a Database Language
- SQL is a non procedural language



Why SQL:

It allows users to

- Access data in Relational Database Management Systems
- Describe the data
- Define and manipulate data
- Embed with other languages
- Create and drop databases and tables
- Create views, procedures and functions
- Set permissions for views, stored procedures and tables

Module# 36: Data Science with SQL

SQL is one of the most requested skills for querying and managing data in Data Science:

- Handling Structured Data
- Data Processing
- Machine Learning

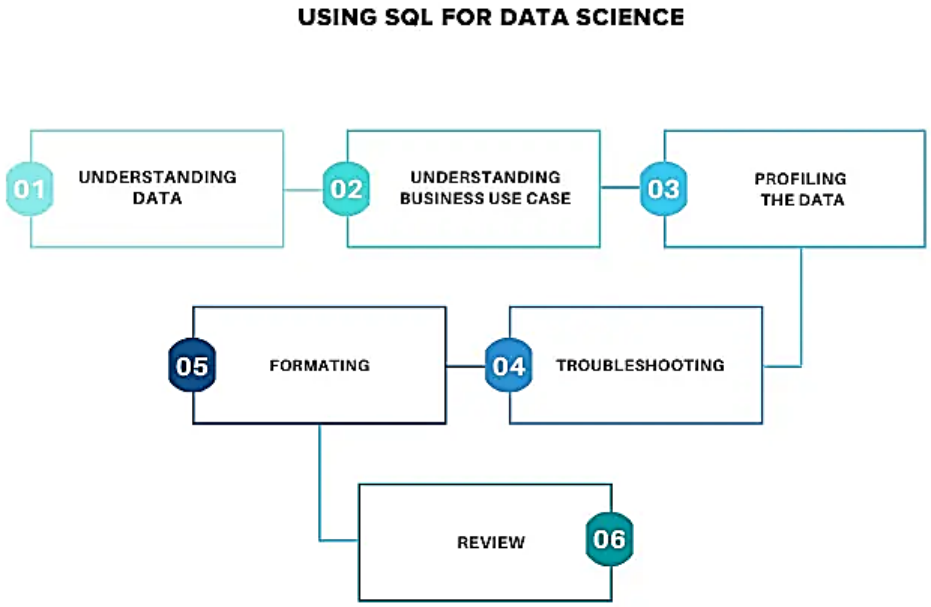
Why SQL is important in Data Science:

- Reliability
- Adaptability
- Scalability
- Predictability
- Manageability

Data Science Involves:

- Extracting, processing and analyzing large data volumes
- Data Scientist needed tools and techniques to manage it

- A large %age of data is available in RDBMS
- SQL is necessary to learn by Data Scientists



Module# 37: SQL Attributes For Data Science

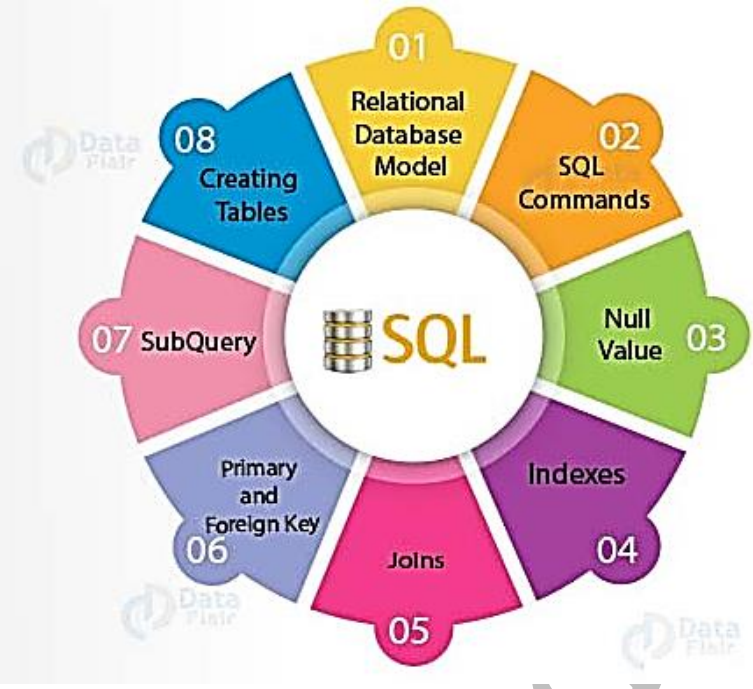
SQL helps data scientists understand their datasets by giving them tools to investigate and visualize the data:

- SQL integrates with Python, R, and other language
- Easier for data scientists to share data with others and to present it clearly
- SQL is Standard API for Relational Databases

When Data is larger and more complex, we need to extract it from the database:

- To Store
- To Query
- For Analytics
- For Visualization

CS442- Introduction of Data Science



Module# 38: SQL Commands

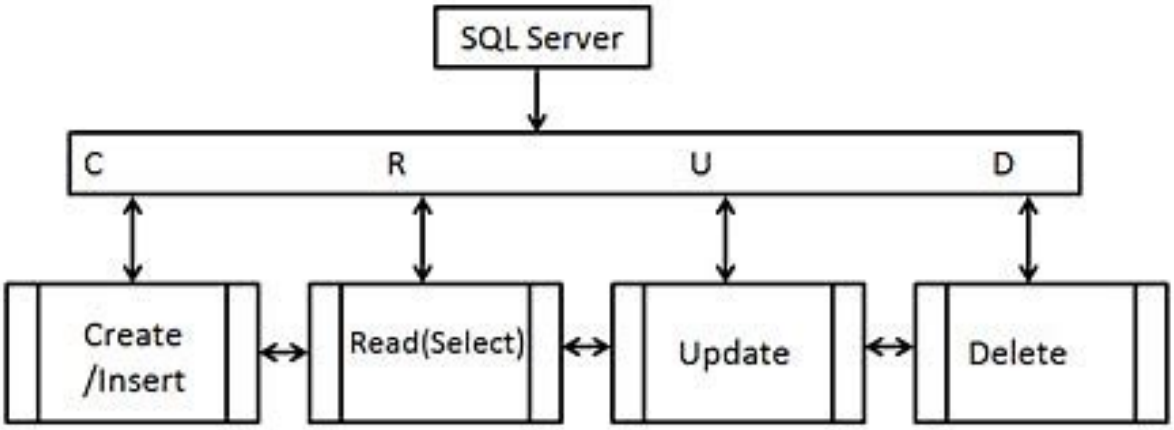
SQL commands are the language inputs to manipulate and handle relational databases:

- To manage and analyze the data in specific methods
- Aimed at simplifying the strenuous process of extracting data
- Acting as a language communicator between the human and the computer system carrying the load

Following are the important SQL Commands essential for Data Science :

- Create Database
- Create Table
- Insert Into
- Select
- Update
- Deleted

- Drop Table

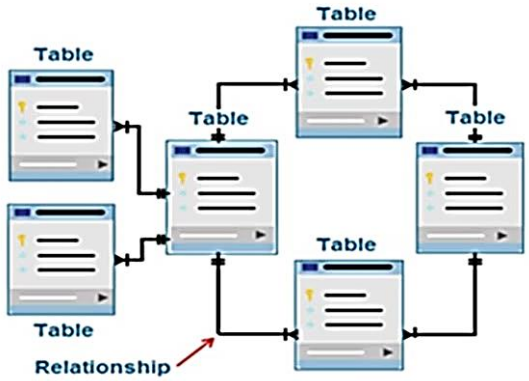


Module# 39: Relational Databases

Relational Databases:

There are many RDBMS in the market to work with:

- Oracle - Licenses
- MS SQL - Licenses
- Microsoft Access - Licenses
- Postgre SQL – Open Source
- MySQL – Open Source



Oracle:

- Owned by Oracle Corporation
- Can Manage Large Data Volumes
- Allows Parallel & Multi Processing
- Supports Major OS:
    - Windows
    - Linux
    - Unix
    - OS/2
- Consistent & Locking Mechanism
- Table Compression
- Data Mining/Data Ware Housing
- Partitioning & Parallel Processing

MS SQL Server:

- Owned by Microsoft
- Supports TSQL, ANSI SQL
- High Performance & Availability
- Database Mirroring & Snapshots
- XML Integration
- Service Broker
- DDL Triggers
- Ranking Functions
- Database Mail

Microsoft Access:

- Entry Level DBMS
- Uses Jet SQL
- Easy to use Graphic Interface
- Support Data Import/Export to many formats
- Support Parametrized Queries
- Can be integrated with:
    - .NET, Visual Studio through ADO, DAO
- File Server Based Database
- Does not support: Stored Procedures, Database Triggers & Transaction Logs

Postgre SQL:

- Open Source
- Object Relational Database

- Reliable & Robust
- High Performance & Secure
- Partitioning & Inheritance
- Backup, Restore & Replication
- Custom Functions, Procedures, Views and DDL
- Indexing & Constraints

MySQL:

- Open Source
- Supports Many Platforms
    - Windows
    - Linux
    - Unix
    - Mac OSX
- Robust SQL Database Server
- High Performance
- High Availability, Scalability Flexibility and Security
- Supports Data Warehouse

CS442- Introduction of Data Science

Module# 40: SQL Data Science Case Study

Sales Forecasting

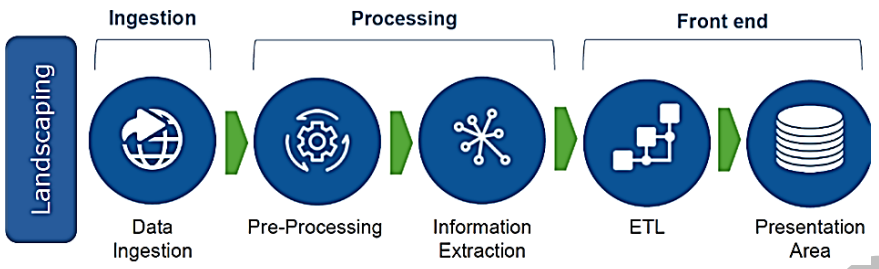
Problem Statement:

- For any departmental store, it is important to have a rough idea of their sales, so that they can plan their inventory accordingly
- Determining the sale of high-selling and low-selling products of each category is also key for inventory planning

CS442- Introduction of Data Science

In this data science project, you will use Python programming language to predict the sales of each department of a store using the Walmart dataset:

Data pipeline



- Walmart Dataset has sales data of 45 stores based on store, department and week
- Size and type of each store has been mentioned
- Holidays weeks have been provided
- Price markdown data (almost like discount data) has been mentioned
- A few macro-indicators like CPI, Unemployment rate, Fuel price etc. are also provided.

Learning Outcomes:

- Exploratory Data Analysis techniques
- Handling Missing values in a dataset
- Using Univariate Analysis
- Performing Bi-variate Analysis
- Time Series models Implementation
- Using multiple metrics to compare the performance of different models

CS442- Introduction of Data Science

Module# 41: What is NoSQL

Not Only SQL data is not split into multiple tables, as it allows all the data in a single data structure. It is suitable when:

- When you work with a huge amount of data
- No need to run the expensive joins
- Highly scalable and reliable and designed to work in a distributed environment

NoSQL Data Types:

- **Document Based**

Store data in JSON objects. Each document has key-value pairs like structures. Flexible and allow to change the structure anytime

- **Key Value Data Bases**

Stores the data as key-value pairs. Here, keys and values can be anything like strings, integers, or even complex objects. They are highly partition-able and are the best in horizontal scaling.

NoSQL Data Types:

- **Wide Column Based**

Stores the data in records similar to any relational database but it has the ability to store very large numbers of dynamic columns

- **Graph Based**

Store data in the form of nodes and edges. The node part of the database stores information about the main entities like people, places, products, etc., and the edges part stores the relationships between them.

CS442- Introduction of Data Science

Module# 42: Data Science With NoSQL

NoSQL in Data Science is becoming more important:

- NoSQL systems are distributed, and designed for large-scale data storage
- Parallel data processing across a large number of commodity servers
- Use non-SQL languages and mechanisms to interact with data
- Support APIs that convert SQL queries to the system‘s native query language or tool

NoSQL systems are designed to scale to thousands or millions of users doing updates as well as reads required by major internet companies, such as:

- Google
- Amazon
- Facebook & More
- Exploratory & Predictive Analytics

Which had challenges in dealing with huge quantities of data which conventional RDBMS solutions could not cope.

NoSQL databases raised in recent years Motivated by requirements of Web 2.0 applications**:**

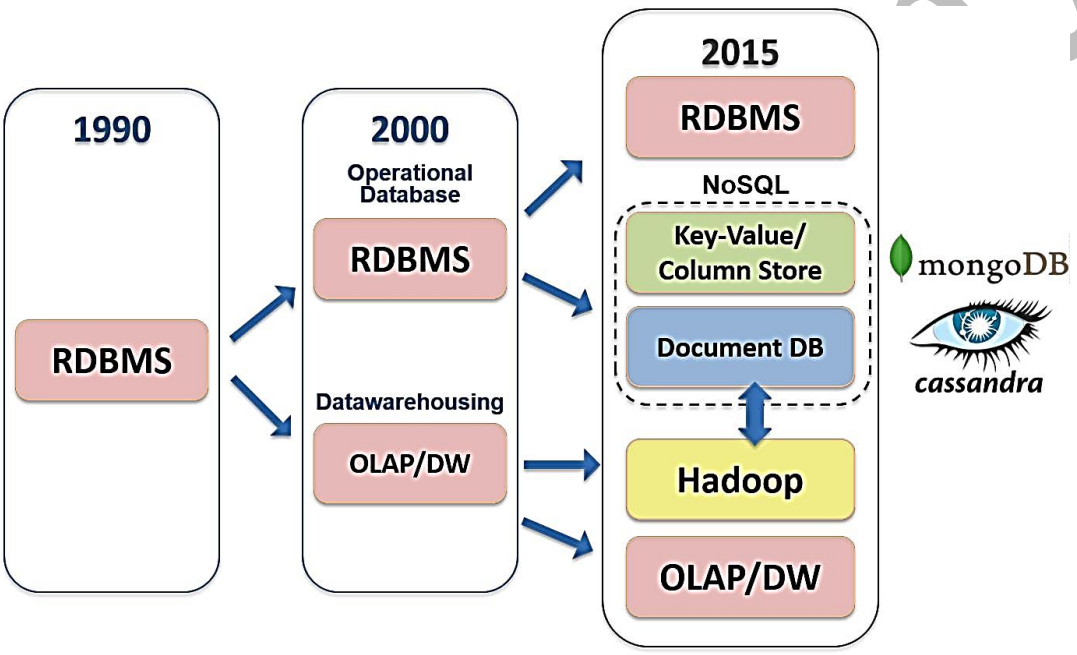
- Strict relational schema is a burden for web applications like blogs
- Databases have to be flexible and agile to support easy schema evaluation
- Exhibit the ability to store and index arbitrarily big data sets while enabling a large amount of concurrent user requests

CS442- Introduction of Data Science

Module# 43: NoSQL Uses in Data Science

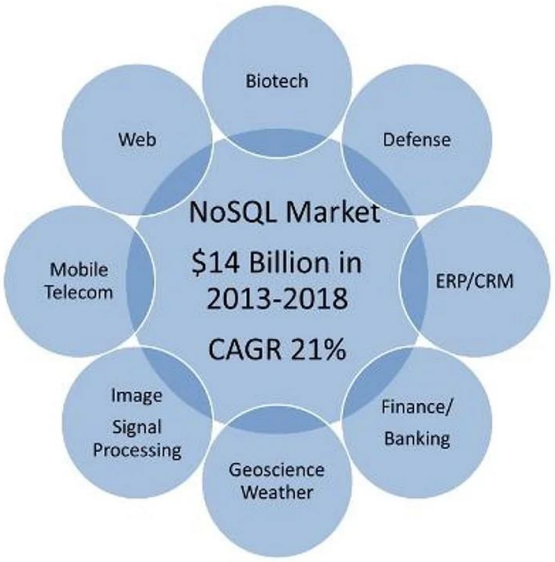
Organizations collecting large amounts of unstructured data are increasingly turning to non-relational databases:

- Focus on analytical processing of large scale datasets
- Increased scalability over commodity hardware
- Ability to store and index arbitrarily big data sets



Major Social Media and Tech and corporate company Applications are using on NoSQL:

- Facebook
- Twitter
- Amazon
- Microsoft
- SAP
- Capgemini
- Qualcomm
- JP Morgan



NoSQL Projects in Data Science:

- Resume Parsing Application

(Natural Language Processing, Python, ML, Elastic Search)

- Building a Chatbot for Customer Care and Support Services

(NLP, Python, ML. Text Files, BangDB)

- Driver Availability System

(Node.js, Redis, Python, ML, MySQL, MangoDB)

Module# 44: NoSQL Databases

Top NoSQL Data Bases are:

- MONGO DB
- Elastic Search
- Cassandra
- Dynamo DB
- HBase

CS442- Introduction of Data Science

MANGO DB:

- Document based
- Store Documents in JSON Objects
- Integrate 100 of data sources
- Great fit for single unified data view
- Expecting 100's of Reads & Writes
- Use to store click stream data
- Suitable for customer behavior Analysis
- Some data may loss in case the server crashes

Elastic Search:

- Open Source NoSQL Database
- Highly scalable and consistent
- Used as an analytical engine
- Store, Search and Analyze large volumes of data
- Supports full text search
- Allows search from fuzzy matching
- Used for Chatbots
- Storing & Analyzing Logs Data

CS442- Introduction of Data Science

Casandra:

- Open source distributed database system (Facebook, Google)
- Widely available & scalable
- Can handle thousands of requests per second
- Manage Petabytes of Data
- Suitable for more write operations like Social Network Websites
- Suitable for Health, weather tracking and time series data

Dynamo DB:

- A key-value pair based distributed database system
- Created by Amazon and is highly scalable, consistent & reliable
- It can easily handle 10 trillion requests per day
- Large simple key value queries
- Suitable for OLTP workloads like banking & booking transactions

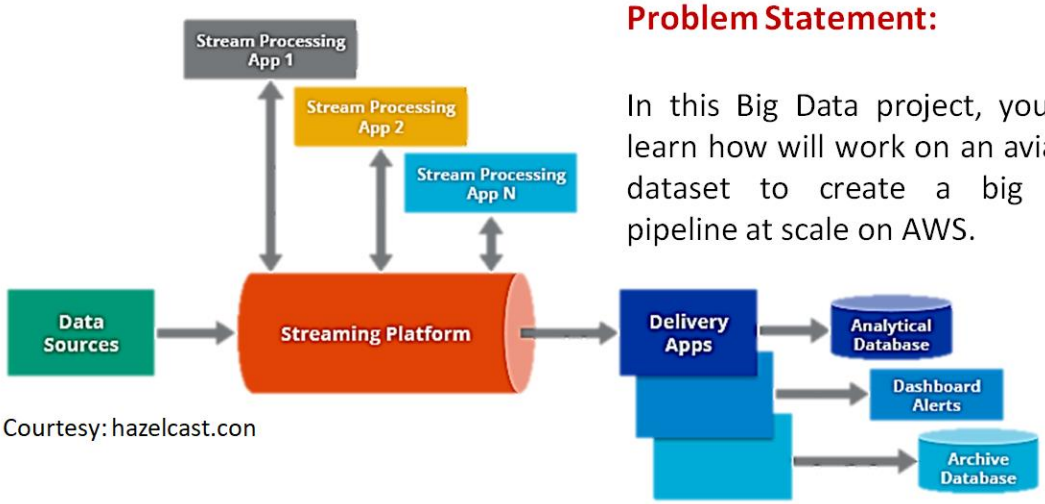
HBase:

- It is also open-source
- Highly scalable distributive database system
- HBase was written in JAVA and runs on top of the Hadoop Distributed File System (HDFS)
- Provides random real time access to the data

- Manage real time messages of millions of people

Module# 45: NoSQL Data Science Case Study

Building Big Data Pipeline



Problem Statement:

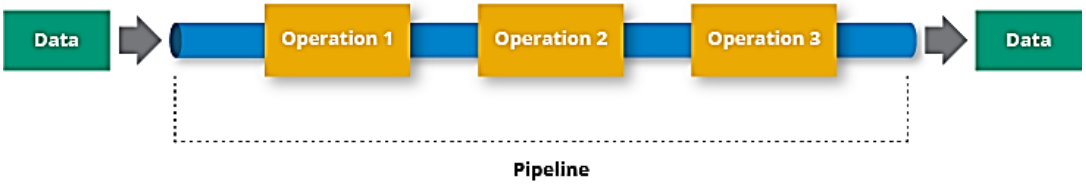
In this Big Data project, you will learn how will work on an aviation dataset to create a big data pipeline at scale on AWS.

In this Big data project you will work with a dataset provided by an API from ICAO. The dataset you will be working with is called Incidents:

- The data format will be JSON

• It has 20+ features which include State of Occurrences, Time, Location, Operator, Model, State of Registry, Flight Phase, Fatalities, etc.

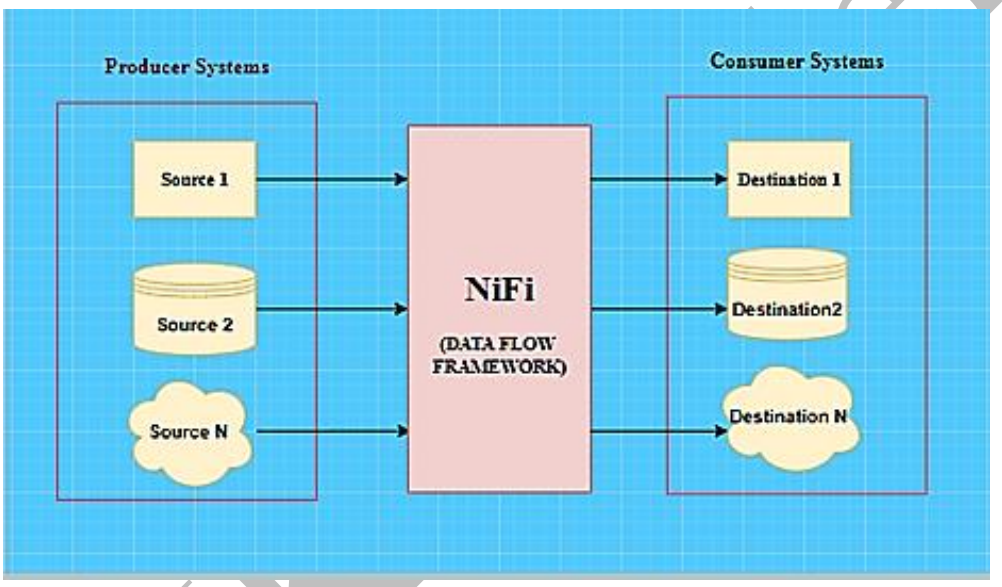
• After hitting the URL of the website, you will be able to access this data



CS442- Introduction of Data Science

Learning Outcomes:

- Introduction to Big Data and Big Data Pipeline
- Explore Data Architecture using Nifi, Kafka, and Hive
- Apache Kafka Vs Apache Flume
- Optimization Techniques in Apache Hive
- Understanding HDFS and Druid
- Transferring Data from Nifi to Kafka
- Using MySQL and AWS Quicksight

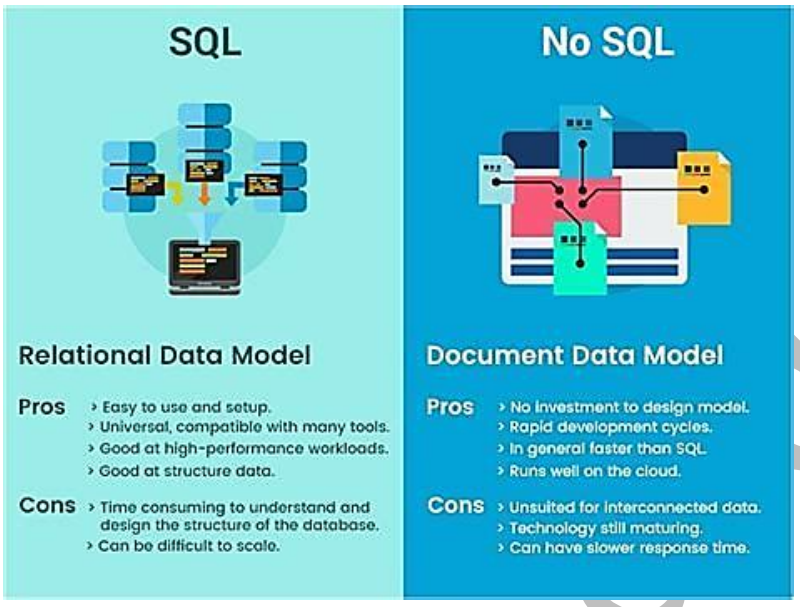


CS442- Introduction of Data Science

Module# 46: SQL Vs NoSQL

The major differences between SQL and NoSQL databases:

This by no means an exhaustive list of differences between the two databases. But hopefully, you got a good overview of the two.



Pros

- Flexible Queries

Enables support for diverse workloads, abstracts data and allows engines to optimize queries to fit on-disk representations

- Reduces Data Storage Footprint

Due to normalization and other optimization opportunities, a reduced footprint maximizes database performance and resource usage

- Strong Data Integrity

Atomicity, consistency, isolation and durability, are database properties that guarantee valid transactions

- **Scalable & Highly Applicable**

Designed to support seamless, online horizontal scalability without significant single points of failure

- **Flexible Data Models**

Do not require developers to make up-front commitments to dynamic data models

- **Dynamic Unstructured Schema**

Documents can be created without a defined structure first, which enables each to have its own unique structure

- **High Performance**

A limited database functionality range enables high performance amongst many NoSQL databases.

Cons

- Rigid Data Models

Requires careful up-front design to ensure adequate performance and resistance to evolution. SQL has a predefined schema, so changing it often includes downtime

- Limited Horizontal Scalability

It is either completely unsupported, supported in an ad-hoc way or only supported on relatively immature technologies

- Single Point of Failure

Non-distributed engines are mitigated by replication and failover techniques.

- Vague Interpretation

Belief that it supports NoSQL systems, ACID interpretations can be too broad to make clear determinations about database semantics

- Distributed System Problems

NoSQL systems, encountering such problems is common and may require SME trouble shooting

- Lack of Flexibility in Access Patterns

The on-disk representation of data leaks into the application's queries and leaves no room for NoSQL engines to optimize queries.

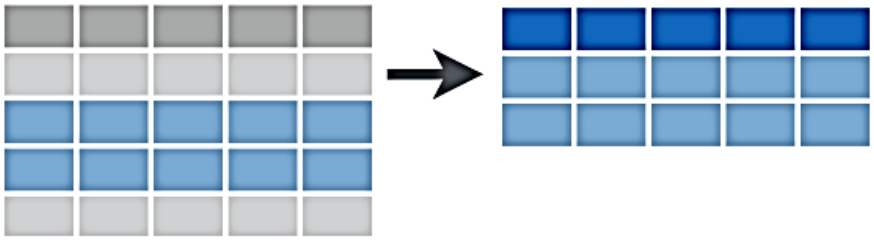
Module# 47: What is Data Wrangling

The Process of Organizing and Completing Raw Data and Standardizing for:

- Easy Access
- Consolidation
- Analysis

It Involves:

- Mapping data fields from source to destination targeting a field, row, or column in a dataset

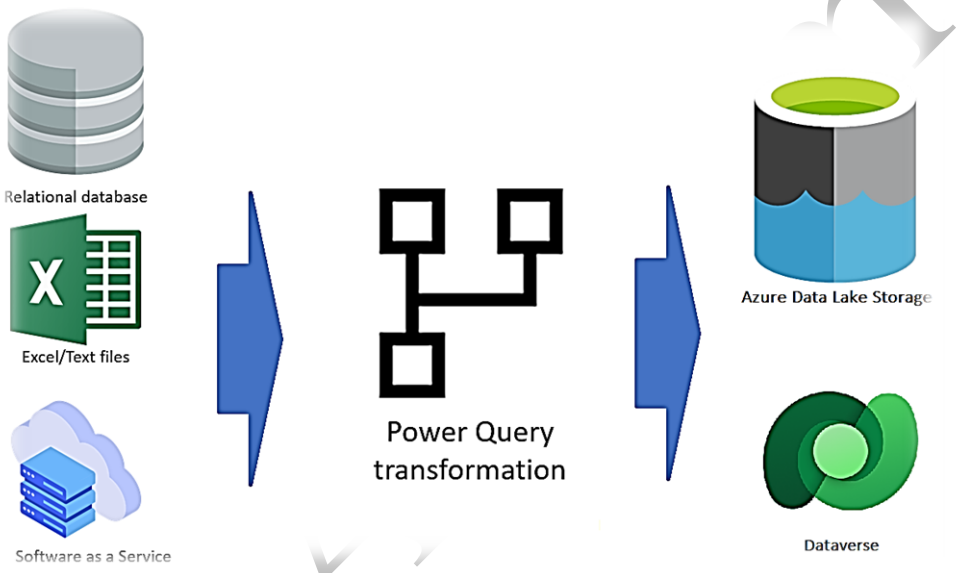


Using the data to process it further for Business intelligence (BI), Reporting and Improving business processes by ensuring that the data is ready for automation.

CNIC	Name	Gender	Date of Birth	District
1730112405734	muhammad saleem khan	Male	8/10/1965	Abbotbad
1310158411045	mushtaq ahmad	Male	11/28/1964	Abbottabad
1330217035737	sardar aurangzeb	Male	3/17/1963	Abbottabad
1310134254423	ahmad faisal	Male		Abbottabad
1210109652315	Nasir khan	Male	3/29/1971	Abbottabad
1310150357095	Almas Khan		3/1/1964	Abbottabad
6110135118523	Mohammad Safder	Male	4/12/1964	Abbottabad
1310114714568	Nighat Yasmeen	Female	5/5/1963	Abbott Abad
1310171457106	Muhammad Amin Khan	Male	9/1/1965	Abbottabad
1310191858889	Tahir Mehmood	Male	June 1956	Abbottabad
1310167188459	Muhammad Irshad	Male	3/13/1964	Abbottabad
1330129740662	Muhammad Saeed	Male	11/15/1961	Abbottabad
1310109880117	Waqas Ahmad Awan	Male	2/19/1967	Abottabad
1310108962142	Iqbal Jan	Male	1/1/1966	Abbottabad

Implementing an action like following to produce the required output:

- Joining
- Parsing
- Cleaning
- Consolidating
- Filtering



Module# 48: Data Wrangling Process

Discovering allows you to:

- Understand your data
- How it's useful for analytic exploration and analysis

Validating:

- Identifies and surfaces data quality and consistency issues

Cleaning:

- Lets you fix and standardize the data that might distort your analysis

Structuring:

- Gives you the ability to format data of all shapes and sizes to work with traditional applications

Enrichment:

- Allows you to take advantage of the wrangling you've already done

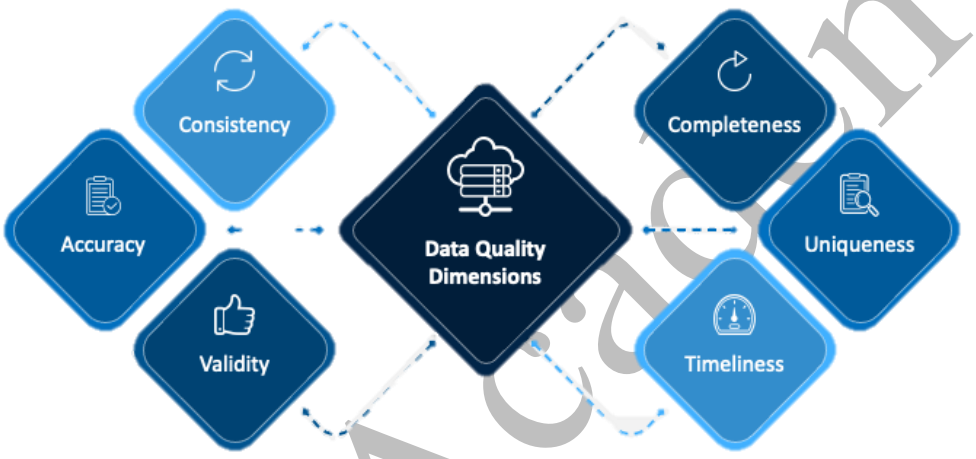
Publishing:

- Provides you the ability to plan for and deliver data for downstream analysis

Module# 49: Data Wrangling Vs Data Cleansing

DATA CLEANING	DATA WRANGLING
Process of removing corrupted or inaccurate records from a table, database or record set.	Process of transforming and mapping data from one raw data into another form with the intent of making it more appropriate and valuable for various task.
Also called Data Cleansing	Also called Data Munging

Module# 50: Data Quality



- **Data Consistency**

Keeping information uniform across the network and between various applications on a computer. There are three types:

- Point in time consistency
- Transaction consistency
- Application consistency

- **Data Completeness**

The percent of all required data currently available in a dataset.

- **Data Accuracy**

Error-free records that can be used as a reliable source of information. It is the first and critical component of the data quality framework

- **Data Validity**

The process ensures the delivery of clean, clear data and checks integrity and validity of data

- **Data Uniqueness**

There should be no data duplicates reported. Each data record should be unique, otherwise the risk of accessing outdated information increases

- **Data Timeliness**

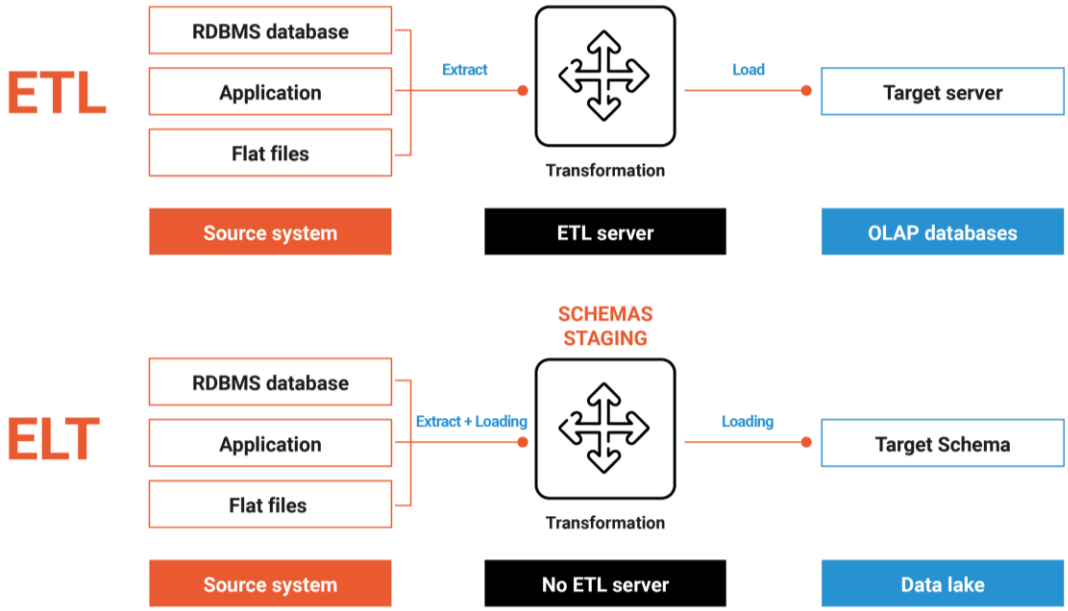
In relation to data quality content, timeliness has been defined as the degree to which data represent reality from the required point in time

Module# 51: ETL Vs ELT

ETL & ELT:

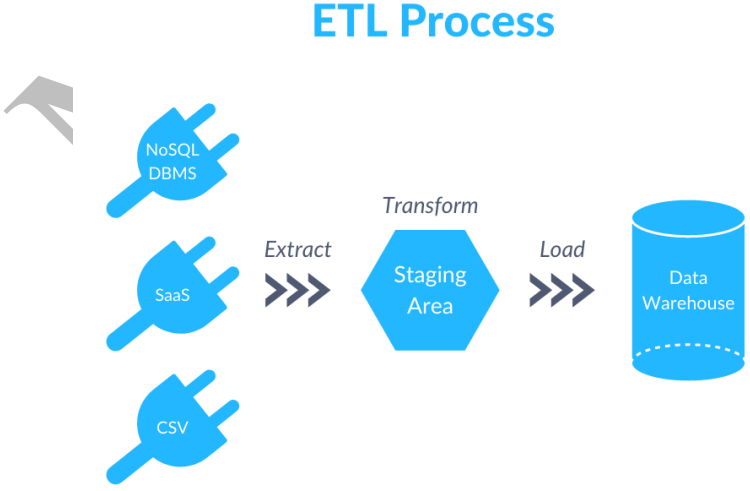
- Both are data processes that allow internal and external business users to access data
- The users can analyze data, make insights, and build models

CS442- Introduction of Data Science



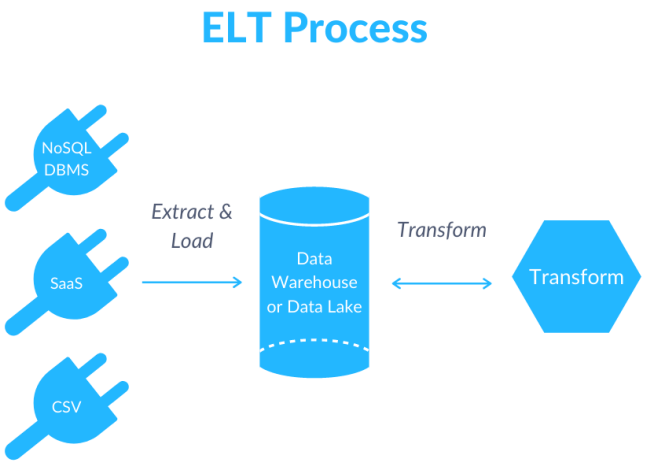
ETL:

- East to manage
- Mature Process
- Best for smaller Data Sets
- Loading time may be slower
- Does not use Data Lake
- May need to change data format



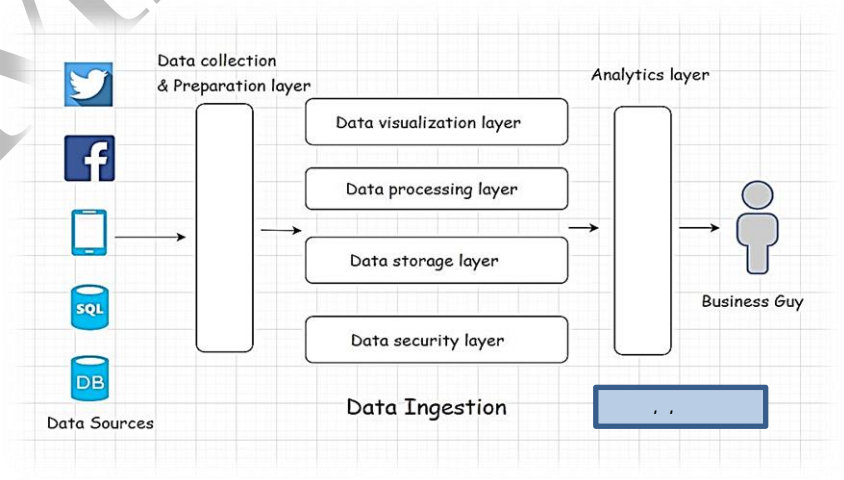
ELT:

- Faster loading from the source
- Easier to automate
- Use Data Lake structure
- Data staging is not required
- TCO may be higher
- Suitable for real time data acquisition and data streaming



The process of obtaining and importing data for immediate use or storage in a database.

- Data can be streamed in real time or ingested in batches
- Faster acquisition and data streaming
- Can be ingested both in Data Warehouse & Data Lake



CS442- Introduction of Data Science

Module# 52: Data Wrangling Tools

Following are some the top Data Wrangling Tools:

- Excel/Spread Sheets
- Talend
- Trifacta
- Tabula
- Google Data Prep
- Data Wrangler

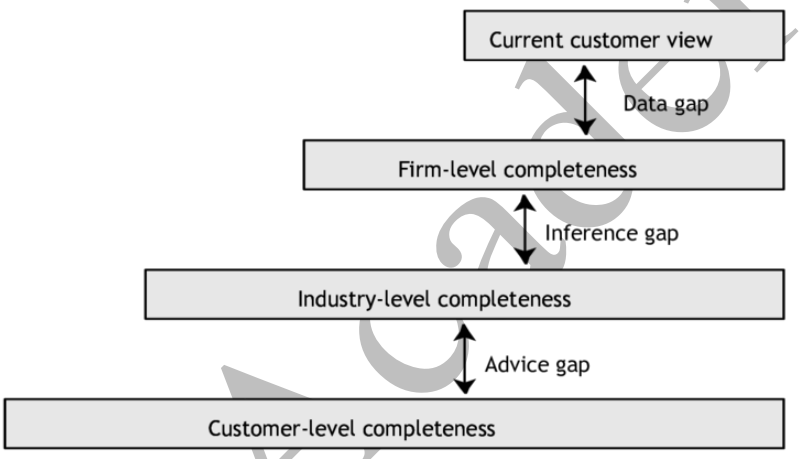
Data Wrangling in machine learning is a huge necessity in recent times because of the huge amounts of data to be processed every day:

- Scripting Languages & Software Development Techniques
- Database Development Technologies & Techniques
- Strong infrastructure of data storage
- Data wrangling tools & Techniques

Module# 53: Data Completeness

The comprehensiveness or wholeness of the data:

- No Gaps
- No Missing Values
- No Missing Records
- No Missing Files

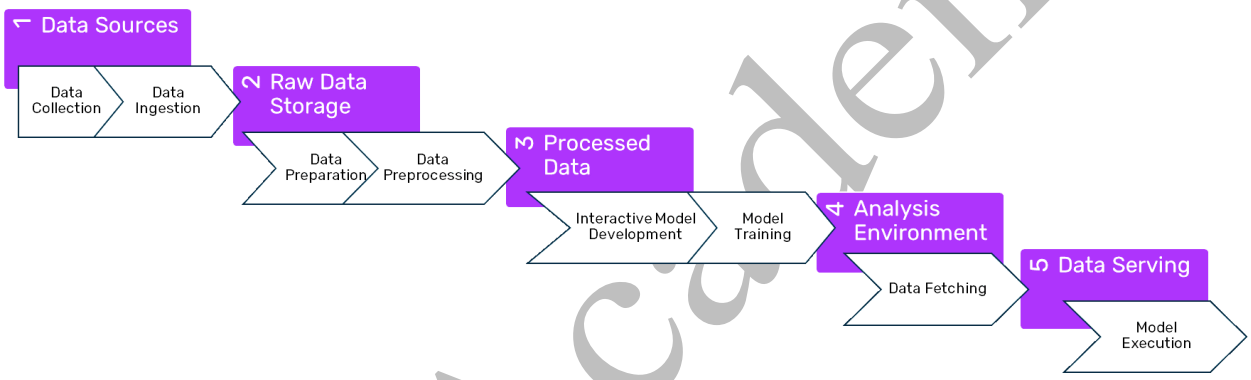


A measure of data completeness is the percentage of missing data entries:

- A column of 500 with 100 missing fields has a completeness degree of 80%
- 20% of missing entries can translate into losing hundreds of thousands of dollars for the bank

Control Input	Data Input
Start transaction	Card details
Cancel transaction	PIN
End transaction	Amount required
Select service	Message

Data Completeness Route in a Big Data Pipeline:



Module# 54: Data Wrangling Example

Using Python

- Data Collected from source system
- Python Data Frame feature
- Data Analyzed by data scientist
- Inconsistencies observed
- Null Values identified
- Average value of the column calculated using python
- Null values replaced with Average value using python

CS442- Introduction of Data Science

In [4]:
```
# Import pandas package
import pandas as pd

# Assign data
data = {'Name': ['Shary', 'Ali', 'Ahmad',
        'Mohsin', 'Shzad', 'Waqar', 'Asad'],
        'Age': [17, 17, 18, 17, 18, 17, 17],
        'Gender': ['M', 'F', 'M', 'M', 'M', 'F', 'F'],
        'Marks': [90, 76, 'NaN', 74, 65, 'NaN', 71]}

# Convert into DataFrame
df = pd.DataFrame(data)

# Display data
df
```

Out[4]:

	Name	Age	Gender	Marks
0	Shary	17	M	90
1	Ali	17	F	76
2	Ahmad	18	M	NaN
3	Mohsin	17	M	74
4	Shzad	18	M	65
5	Waqar	17	F	NaN

In [5]:
```
# Compute average
c = avg = 0
for ele in df['Marks']:
    if str(ele).isnumeric():
        c += 1
        avg += ele
avg /= c

# Replace missing values
df = df.replace(to_replace="NaN",
                value=avg)

# Display data
df
```

Out[5]:

	Name	Age	Gender	Marks
0	Shary	17	M	90.0
1	Ali	17	F	76.0
2	Ahmad	18	M	75.2
3	Mohsin	17	M	74.0
4	Shzad	18	M	65.0
5	Waqar	17	F	75.2
6	Asad	17	F	71.0

CS442- Introduction of Data Science

Module# 55: Data Wrangling Example

Using Excel

Excel Provides Rich Functions to perform various Data Warngling Functions:

- Max, Min
- Trim
- Left & Right Selection
- Conversion
- Lower & Upper Case
- Mid String

MAX

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	21
1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	
1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	
1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	
1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	
1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	
1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	
1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	
1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	

MIN

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	13
1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	
1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	
1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	
1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	
1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	
1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	
1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	
1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	

TRIM

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	Mardan
1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	Karak
1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	D.I. Khan
1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	Mansehra
1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	Peshawar
1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	Chitral Lower
1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	Mansehra
1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	Swabi
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	Abbottabad
1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	Peshawar

CS442- Introduction of Data Science

LEFT

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	Ma
1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	Ka
1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	D.
1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	Ma
1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	Pe
1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	Ch
1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	Ma
1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	Sw
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	Ab
1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	Pe

NUMBER

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	3005779597
1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	3005972506
1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	3005904798
1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	3219806506
1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	3339119465
1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	3025514323
1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	3219981001
1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	3459294157
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	342111883
1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	3018933573

MID

1. CNIC	2. Employee Name	3. Father/Husband Name	4. Gender	5. Date of Birth	6. Domicile	7. Contact Number	Scale	Result
1610127164591	Dr Arshad Ahmad	abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	shad
1730113168921	Dr Muhammad Rasool Jan	sakhi marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	hamm
1730125104461	Dr Fahim Hussain Khan	inayatullah khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	him
1320207468199	Dr SaifUllah Khalid	ghulam rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	ifUl
1730112874135	Dr Arshad Amir	umar amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	shad
1520128355749	Shahzada Muhammad Haider ul Mulk	shahzada muhammad mutaul mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	ada
1350390804545	Dr Niaz Muhammad	khalil ur rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	az M
1620213295339	Asghar Khan	mohabat khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	r Kh
1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	mad
1730112845371	Muhammad Naseem	aya khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	mad

Module# 56: Data Wrangling Example

Using Database

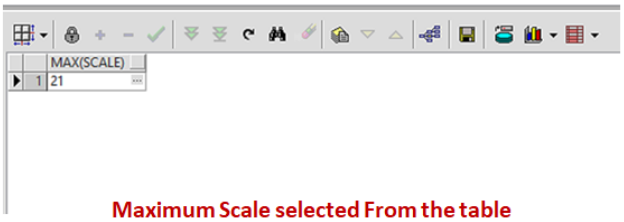
```
SELECT * FROM BPS_DATA B

select max(scale) from BPS_DATA;
```

Employee Data Imported in SQL Database from the Source System:

Export Query Results...

	CNIC	EMPLOYEE_NAME	FATHER_HUSBAND_...	GENDER	DATE_OF_BIRTH	DOMICILE	CONTACT_NUMBER	SCALE	RESULT
1	1610127164591	Dr Arshad Ahmad	Abdurrahman	Male	9/11/1961	Mardan	3,005,779,597.000	20	
2	1730113168921	Dr Muhammad Rasool Jan	Sakhi Marjan jan	Male	4/20/1963	Karak	3,005,972,506.000	15	
3	1730125104461	Dr Fahim Hussain Khan	Inayatullah Khan	Male	2/22/1964	D.I. Khan	3,005,904,798.000	20	
4	1320207468199	Dr SaifUllah Khalid	Ghulam Rasool	Male	5/28/1962	Mansehra	3,219,806,506.000	15	
5	1730112874135	Dr Arshad Amir	Umar Amir	Male	11/9/1961	Peshawar	3,339,119,465.000	13	
6	1520128355749	Shahzada Muhammad Haider ul Mulk	Shahzada Muhammad Mutaul Mulk	Male	4/1/1968	Chitral Lower	3,025,514,323.000	21	
7	1350390804545	Dr Niaz Muhammad	khalil ur Rehman	Male	2/27/1962	Mansehra	3,219,981,001.000	15	
8	1620213295339	Asghar Khan	Mohabat Khan	Male	4/14/1973	Swabi	3,459,294,157.000	19	
9	1730112405734	muhammad saleem khan	hakim khan	Male	8/10/1965	Abbottabad	342,111,883.000	18	
10	1730112845371	Muhammad Naseem	Aya Khan	Male	6/11/1966	Peshawar	3,018,933,573.000	18	

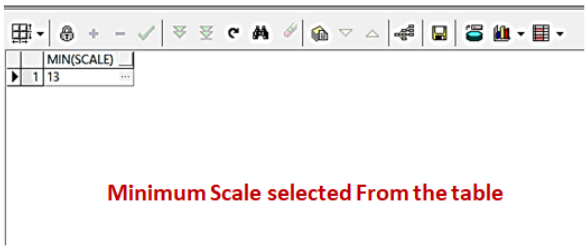
```
select max(scale) from BPS_DATA;
```



	MAX(SCALE)
1	21

Maximum Scale selected From the table

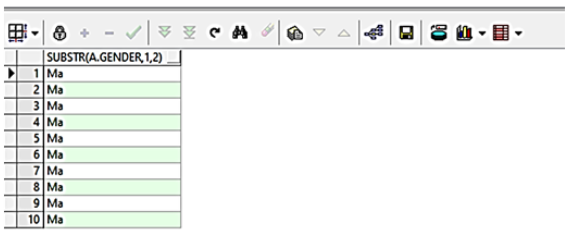
```
select min(scale) from BPS_DATA;
```



	MIN(SCALE)
1	13

Minimum Scale selected From the table

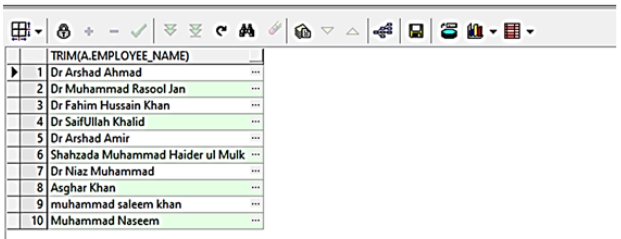
```
select substr(a.gender,1,2) from BPS_DATA a;
```



	SUBSTR(A.GENDER,1,2)
1	Ma
2	Ma
3	Ma
4	Ma
5	Ma
6	Ma
7	Ma
8	Ma
9	Ma
10	Ma

Selected Substring (First 2 Characters) from Gender Column

```
select trim(a.employee_name) from BPS_DATA a;
```



	TRIM(A.EMPLOYEE_NAME)
1	Dr Arshad Ahmad
2	Dr Muhammad Rasool Jan
3	Dr Fahim Hussain Khan
4	Dr SaifUllah Khalid
5	Dr Arshad Amir
6	Shahzada Muhammad Haider ul Mulk
7	Dr Niaz Muhammad
8	Asghar Khan
9	muhammad saleem khan
10	Muhammad Naseem

Used Trim Function (Removed Extra Spaces) from Name Column

Module# 57: What is Python

Python is a great object-oriented, procedural, interpreted, and interactive programming language:

- Python is powerful

- Python is fast
- Plays well with others
- Runs everywhere
- Friendly & easy to learn

Python combines remarkable power with very clear syntax:

- It has modules, classes, exceptions
- Less number of code lines
- Very high level dynamic data types
- Interfaces to many system
- Rich and growing libraries

New built-in Python modules are easily written in C or C++ or other languages, depending on the chosen implementation:

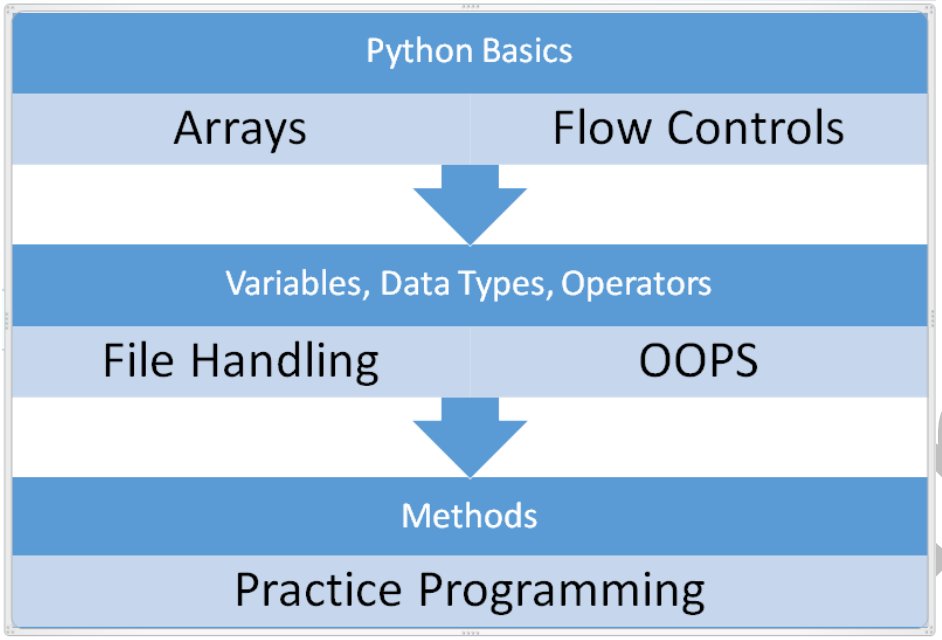
- Usable as an extension language for applications written in other languages
- Need easy-to-use scripting or automation interfaces

The Python Community:

- Vast, Diverse
- Growing
- Open Source
- Python Wiki
- Python Software Foundation

Module# 58: Python Composition

CS442- Introduction of Data Science



There are six basic data types used in python programming:

- Numbers
    - Integers
    - Floats
- Strings
- Lists
- Dictionary
- Sets
- Tuples

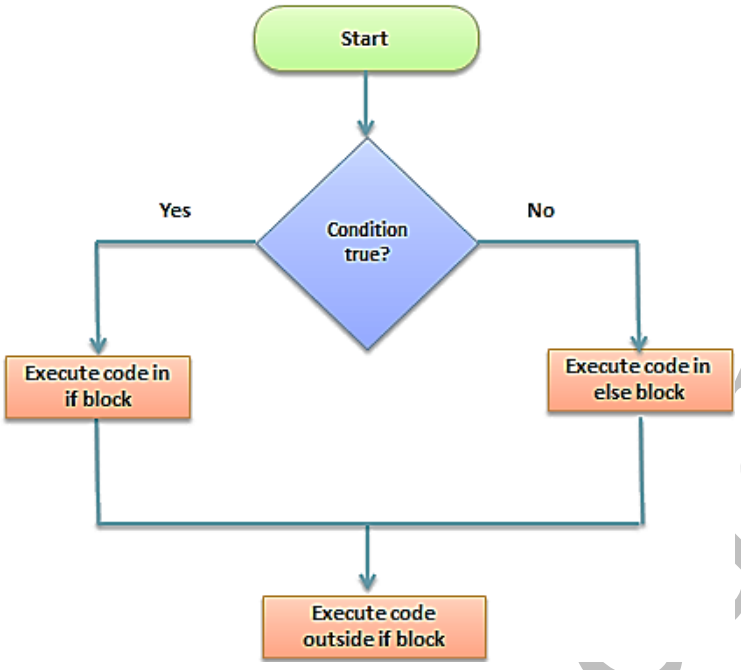
Conditional statements in python programming language:

- **True**
- **False**

Conditional statements in python programming language:

- If

- If Else
- If Elif
- Nested If



Naming Rules:

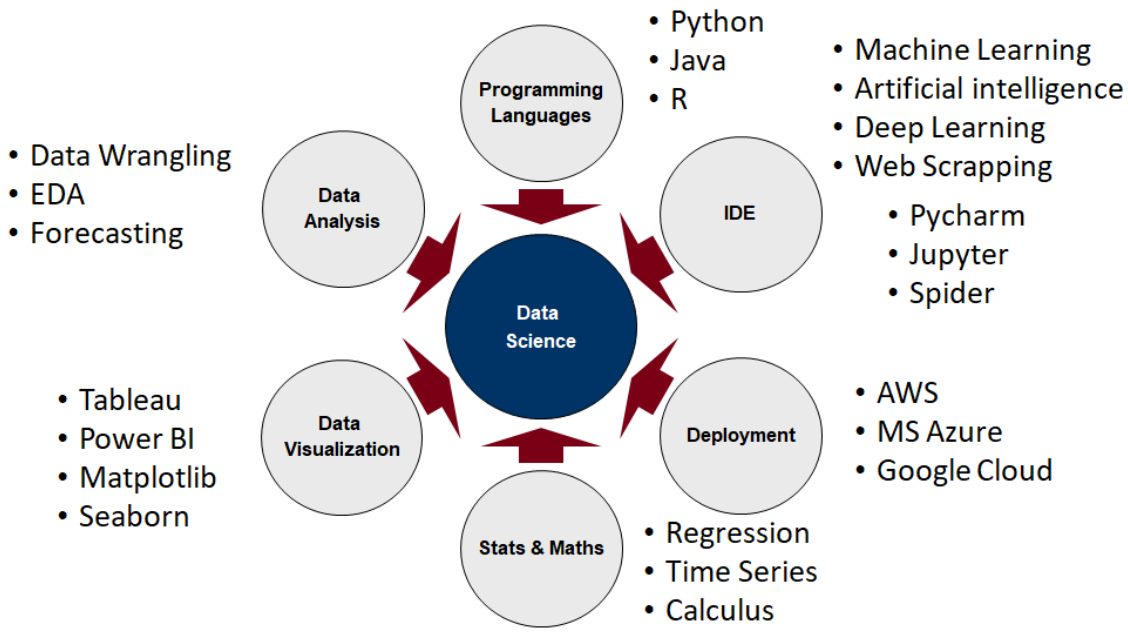
- Names are case sensitive
- Can not start with a number

Reserve Words:

and, assert, break, class, continue, def, del, elif, else, except, exec, finally, for, from, global, if, import, in, is, lambda, not, or, pass, print, raise, return, try, while

Module# 59: Data Science Environment

CS442- Introduction of Data Science



Data Science Environment Setup

1. Python
2. Terminal
3. Code Editor
4. Virtual Environment
5. Version Control
6. Notebooks

Module# 60: Python: Python Vs R Programming

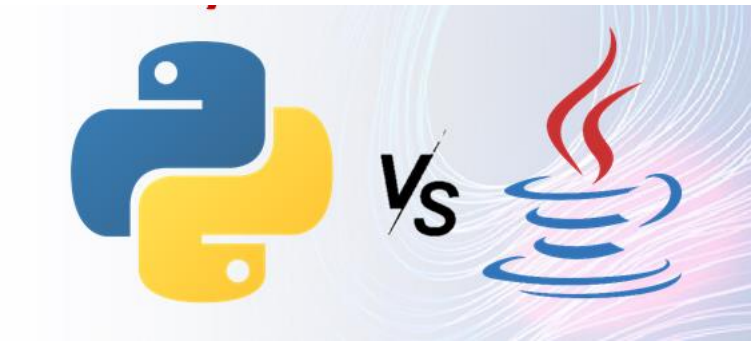
CS442- Introduction of Data Science

General	Python is a general-purpose programming lanuage for data analysis and scientific computing.	R is a functional programming enviornment and language for statstical computing and graphics.
Objective	Data Science, Web Developoment, Embedded Systems	Data Science & Statistical Modeling
IDE	**iPython, Pycharm, Jupyter Notebook, Spyder**	**Rstudio, R GUI, R KWARD**
Data Collection	Supports CSV files, **SQL**, **JSON**, and webscraping with **BeautifulSoup.**	Can also import csv files with built-in **readr** library. R's library **RCurl** provides a simple way to make API requests, similar to Python's **requests** package.
Data Analysis	Orgnaize dataframes with **Pandas** filtering, sorting. Python takes a more streamlined approach for data science projects.	Complex data visualizaiton tools make the exploratory data analysis (EDA) process much more complex than Python.

Essential Packages & Libraries	**Numpy, Pandas, matplotlib, scipy, scikit-learn, TensorFlow**	**caret, stringr, ggplot2, knitr, tidyverse, markdown, shiny, forcats, haven**
Database Handling Capacity	Can easily handle large data because there are less constraints for memory usage	R computes everything in memory, so its capabilities are limited by RAM size. A major downfall of R is the inability to handle massive amounts of data
Data Visualization	Despite the capabilities of data visualization tools like **Matplotlib** and **Seaborn**, Python fails to measure up to data visualization features of R.	Developed by and for statisticians, R has complex data visualzatioon features.
Syntax	The 'zen of python' is that there's a proper way to write code.	R doesn't have this set of rules. Also indexing starts at 1, which can be considered unconventional for general programmers.
Learning Curve	Simple and readable code structure makes it easier for beginners to learn. It also allows for object-oriented programing. It also offers a wide range of data structures that you wouldn't expect from a general-purpose language.	R's functional syntax isn't easy for beginners, but not too challenging for those well versed in programming. It also offers a few data structures, but fails to handle large amounts of data.

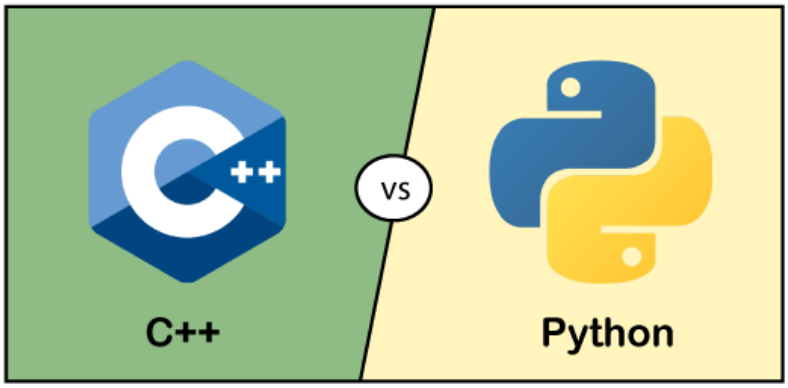
CS442- Introduction of Data Science

Module# 61: Python Vs JAVA

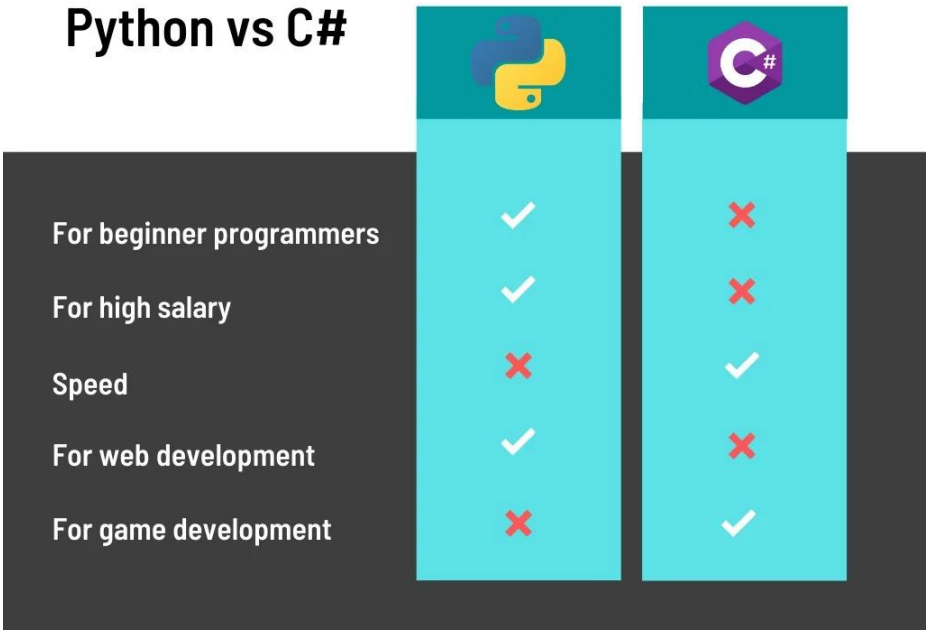


Java		Python
It is the most fundamental language for multiple platforms.	USABILITY	Python has a more high-level programming language
Java is statically typed programming, making it faster.	SPEED	Python, on the other hand, is an interpreter is manually typed, makes it slower.
Java is object-oriented programming language.	LANGUAGE	Python is object-oriented with the advantage of scripting language.
Java Legacy systems are typically larger and numerous.	LEGACY	It has less legacy problem making it difficult for the organizations to copy and paste.
Longer line of code than Python	CODE	Shorter lines of code than Java.
Java database connectivity is popular and widely used to connect.	DATABASES	Pythons access layers are weaker than Java.
Less growth than Python.	SEARCH RESULT	There is significant growth in the search of Python.
The syntax of Java is complex than Python.	SYNTAX	The syntax of Python is easier than Java.
Java because of its statically typed programming is popular for mobile and web applications.	PRACTICAL AGILITY	Python on the other is popular with morerecent choices like ML, AI, Data science, etc.

Module# 62: Python Vs C++



- Python is an interpreted language and it runs through an interpreter during compilation
- C++ is a pre-compiled programming language and doesn't need any interpreter during compilation.



CS442- Introduction of Data Science

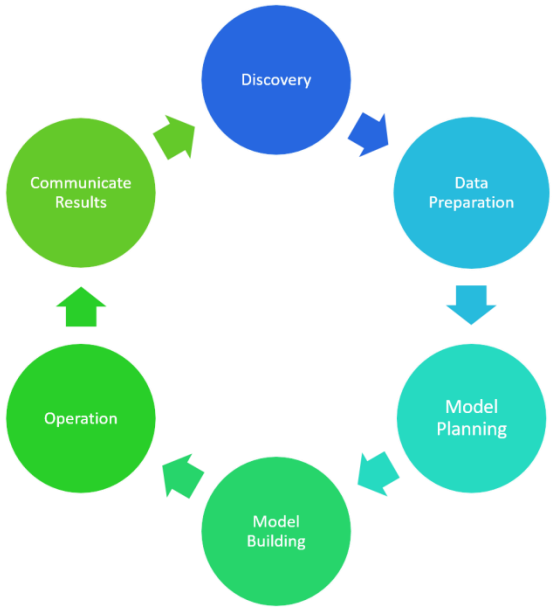
Python VS C++	
#1 Advantages: ▪ Easy to Learn ▪ Easy to access libraries ▪ Scientific community sharing (open source, many libraries) Disadvantages: ▪ Slow ▪ Interpreted (dependencies) ▪ Not typed (errors at runtime)	Advantages: ▪ Execution speed ▪ Pre-Compiled (exe on machine) ▪ Typed (well defined) ▪ Modern professional libraries Disadvantages: ▪ Learning curve ▪ Harder to access libraries (less sharing than in Python)

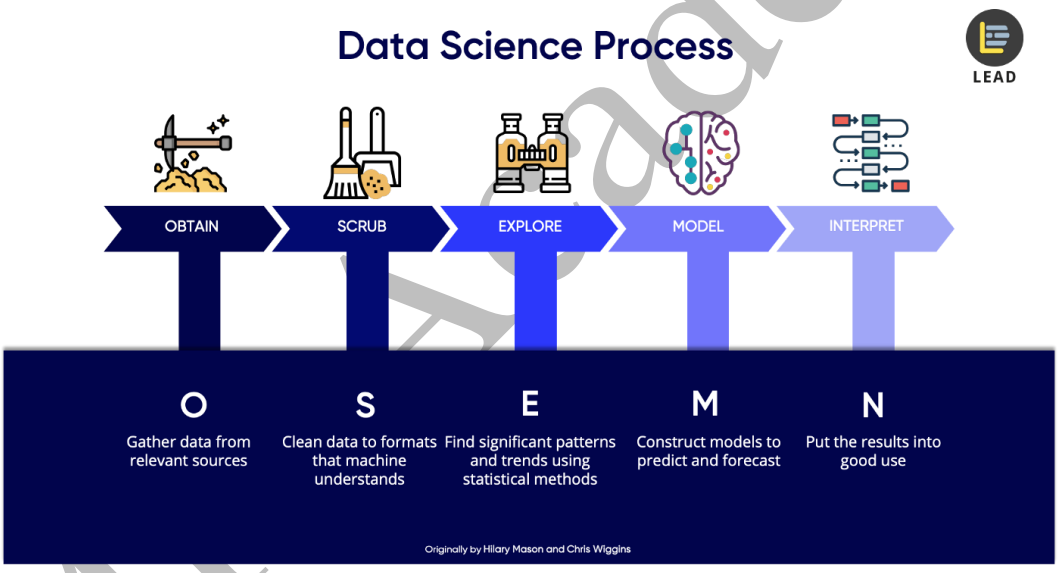
Module# 63: Data Science Process

Data Science Project Lifecycle is divided into five steps:

- Discovery
- Data Preparation
- Model Planning
- Model Building
- Operations
- Communicate Results

CS442- Introduction of Data Science





Module# 64: Data Analysis Using Python

CS442- Introduction of Data Science

Python is more than enough as a programming language if you want to get into machine learning:

- ML Algorithms
- Database Management Languages (SQL)
- Mathematics and Statistics

Following are the Data Analytics steps for Python:

- Import data sets (Pandas)
- Clean and prepare data for analysis (Scikit Learn)
- Manipulate (Pandas DataFrame)
- Summarize data (Pandas)
- Build machine learning models (Keras, Pytorch)
- Build data pipelines

Module# 65: Python Integrated Development Environment

The An Integrated Development Environment (IDE) is an application which facilitates the application development:

- It is a graphical user interface (GUI)-based workbench
- Designed to help developers in building software applications
- It is a combined environment with all the required tools

IDE is a coding tool which allows you to:

- Write, test, and debug your code
- Offer code completion and code insight by highlighting, resource management
- Include debugging tools in an easier way

Why IDEs :

- You can't execute your program in a text editor

- Use on GUI rather than using two different programs
- No plug-ins required
- Auto documentation
- Framework to support complete SDLC

Module# 66: Setting Up Python Development Environment

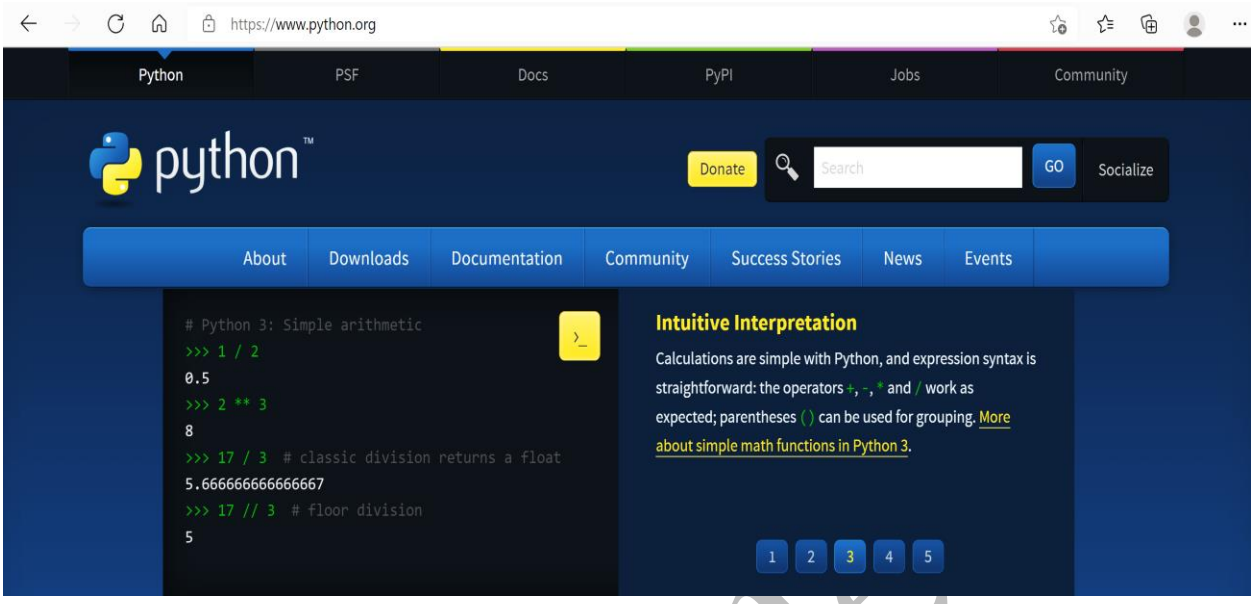
Following are the steps in setting-up Python Development Environment:

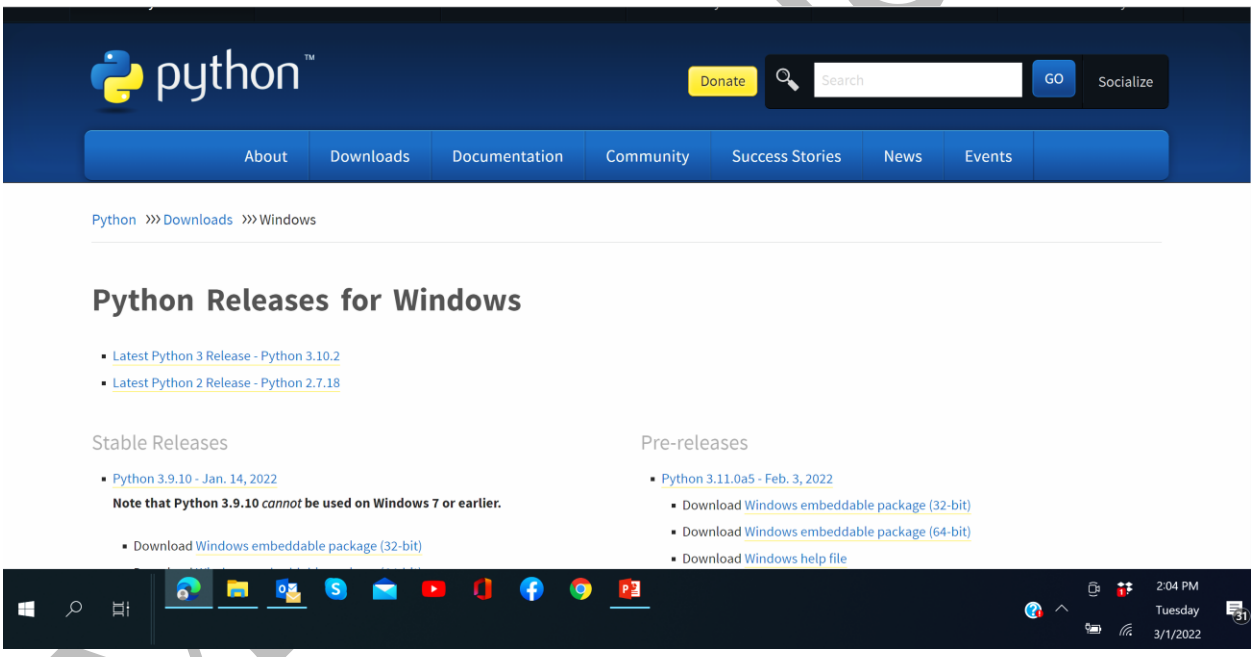
- Numerical
- Computational
- Visualization

What You"ll Learn

1. Where to download Python
2. How to install Python
3. How to open IDLE

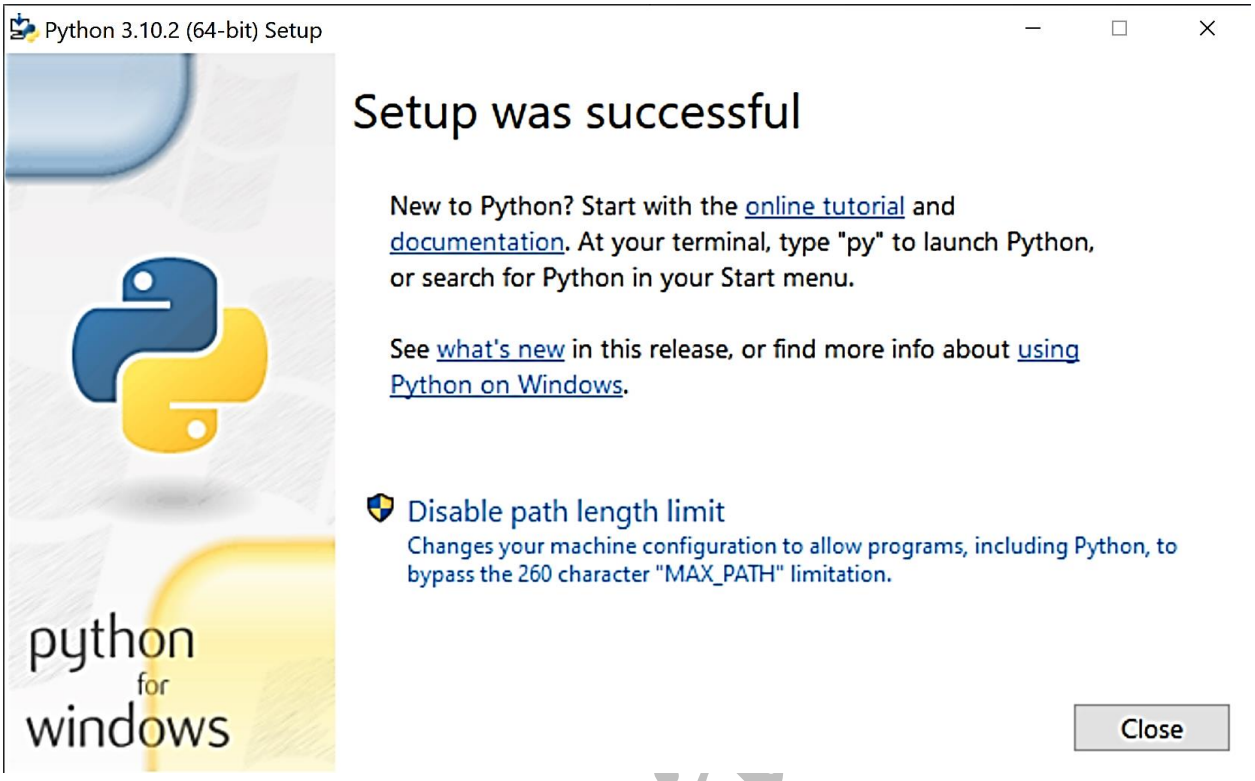
CS442- Introduction of Data Science

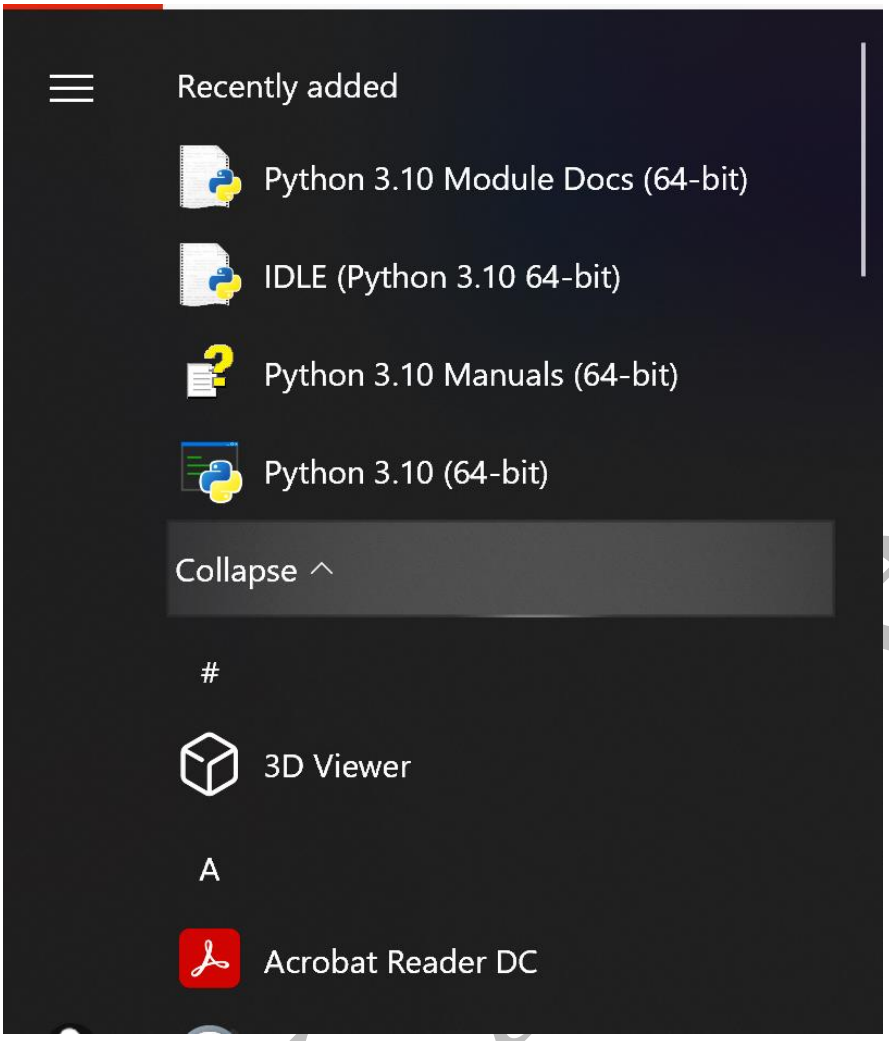




CS442- Introduction of Data Science

CS442- Introduction of Data Science





Module# 67: Python Libraries

There are three types of Python Libraries:

- Numerical
- Computational
- Visualization

Numerical Python Libraries assures calculations with matrices and arrays:

- SciPy
- Pandas
- IPython

- Numpy
- Natural Language Toolkit

Computational Python Libraries:

- Pytorch
- Tensorflow
- Scikit Learn
- Theanos
- Keras

Visualization Python Libraries:

- Matplotlib
- Seaborn
- Plotly
- Altair
- Light GBM

Module# 68: SciPy Python Library

SciPy Python Library is used for Scientific and Technical Computing:

- The basic data structure used by SciPy is a multidimensional array provided by the NumPy module

SciPy Python Library is used for Scientific and Technical Computing:

- SciPy contains modules for:

- Optimization
- Linear Algebra
- Integration
- Interpolation
- Special Functions
- Other common tasks required in science and engineering

Available sub-packages includes:

- **cluster**: hierarchical clustering, vector quantization, K-means
- **constants**: physical constants and conversion factors
- **fft**: Discrete Fourier Transform algorithms
- **fftpack**: Legacy interface for Discrete Fourier Transforms
- **integrate**: numerical integration routines
- **interpolate**: interpolation tools
- **io**: data input and output
- **linalg**: linear algebra routines
- **misc**: miscellaneous utilities (e.g. example images)
- **ndimage**: various functions for multi-dimensional image processing
- **ODR:** orthogonal distance regression classes and algorithms
- **optimize**: optimization algorithms including linear programming
- **signal**: signal processing tools
- **sparse**: sparse matrices and related algorithms
- **spatial**: algorithms for spatial structures such as k-d trees, nearest neighbors, Convex hulls, etc.
- **special**: special functions
- **stats**: statistical functions

- **weave**: tool for writing C/C++ code as Python multiline strings

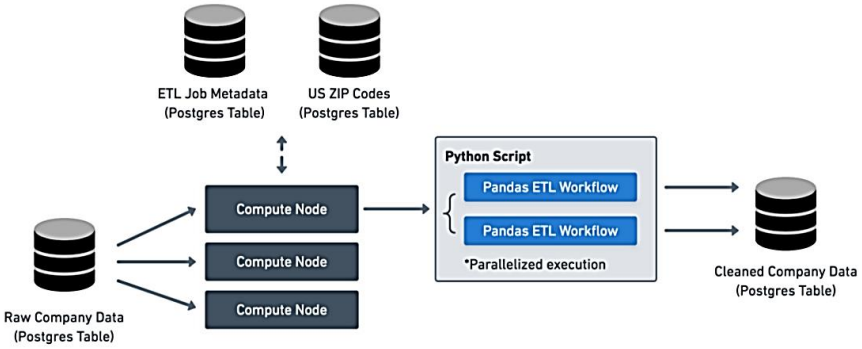
Module# 69: Pandas Python Library

Pandas Python Library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data:

- Data cleansing
- Data fill
- Data normalization

Few More Pandas Python Library Features:

- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data
- And much more



CS442- Introduction of Data Science

Module# 70: IPython Library

IPython (Interactive Python) is a command shell for interactive computing in multiple programming languages:

- Introspection
- Rich Media
- Shell Syntax
- Tab completion
- History

IPython Notebook is web-based graphical interface to IPython combines:

- Code
- Text
- Mathematical Expressions
- Inline Plots
- Interactive Figures
- Widgets

IPython architecture contributes to parallel and distributed computing & facilitates :

- Customer user defined prototypes
- Task Parallelism
- Data Parallelism
- Message cursory using M.P.I
- Multiple programs, multiple data (MIMD) parallelism
- A single program, multiple data (SPMD) parallelism

CS442- Introduction of Data Science

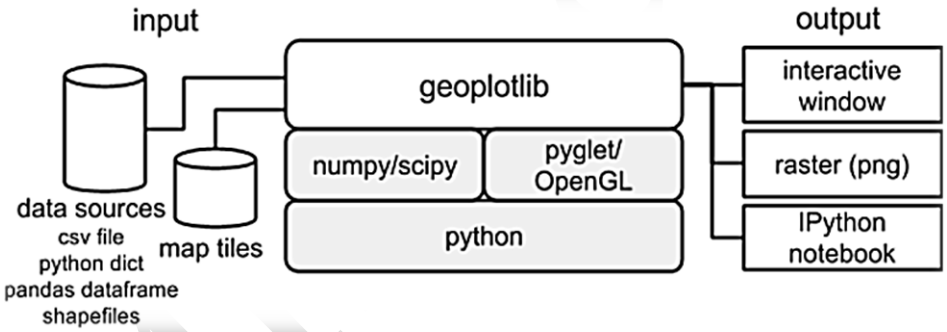
Module# 71: NumPy Python Library

Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing arrays:

- Mathematical Operations
- Logical Operations
- Computational Power

Numerical Python, may be applied in variety of situations:

- Scientific Domains
- Array Libraries
- Data Science
- Machine Learning
- Data Visualization



Module# 72: NLP Python Tool Kit

Natural language processing (NLP) is a field that focuses on making natural human language usable by computer programs to interpret:

- Text Data
- Unstructured Data

CS442- Introduction of Data Science

NLP Employees computational Techniques for purpose of:

- Learning
- Understanding
- Producing Content
- Analysis
- Visualizations

Natural language processing Features & Functions:

- Tokenizing
- Filtering Stop Words
- Stemming
- Tagging Parts of Speech
- Lemmatizing
- Chunking
- Chinking

Natural language processing Features & Functions:

- Using Named Entity Recognition (NER)
- Getting Text to Analyze
- Using a Concordance
- Making a Dispersion Plot
- Making a Frequency Distribution
- Finding Collocations

CS442- Introduction of Data Science

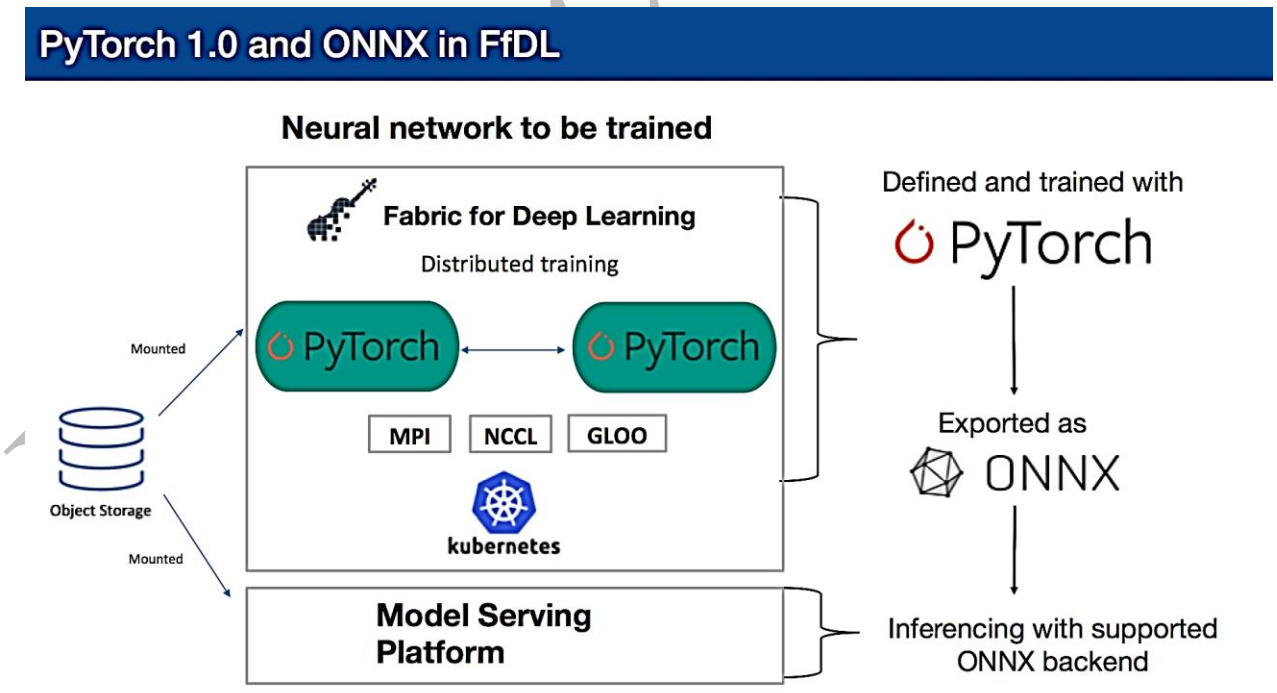
Module# 73: Pytorch Python Library

PyTorch provides two high level features:

- Tensor computation (like NumPy) with strong GPU acceleration
- Deep neural networks built on a tape-based autograd system

PyTorch is defined as an open source machine learning library for Python:

- Free & Open Source
- Provides C++ Interface
- Used by Facebook & Uber
- Deep Learning
- Research & Development
- Easy to use APIs



CS442- Introduction of Data Science

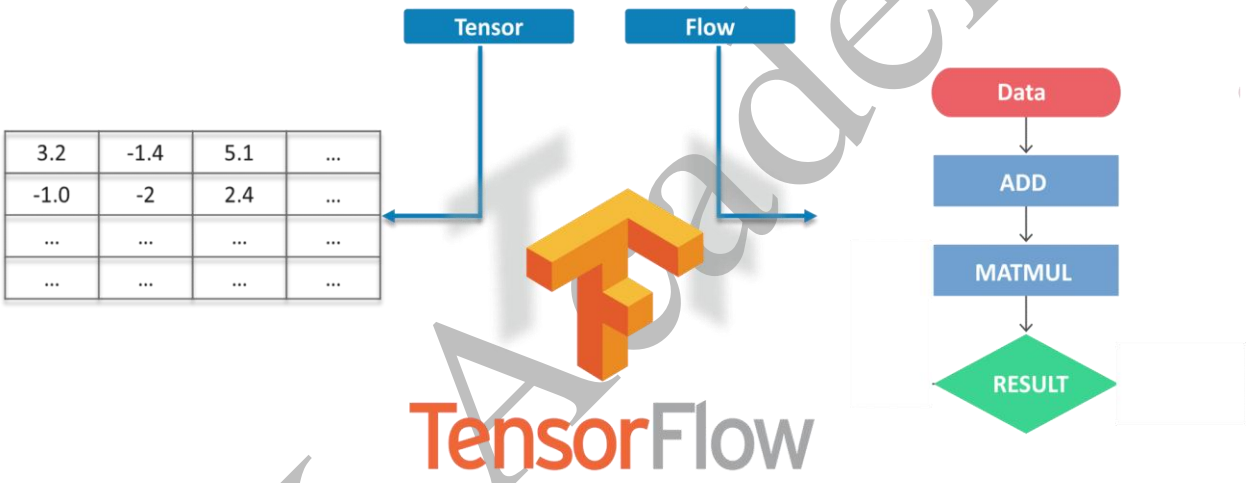
Module# 74: TensorFlow Python Library

TensorFlow is a comprehensive, flexible ecosystem of tools, libraries and community resources:

- It lets researchers push the state-of-the-art in ML
- Easily build models and deploy ML powered application

TensorFlow is an end-to-end open source platform for machine learning:

- It has a comprehensive, flexible ecosystem
- Provides tools, libraries and community resources



Framework for dataflow programming across a range of tasks. Nodes in the graph represent mathematical operations, while the graph edges represent the multi-dimensional data arrays (**tensors**) communicated between them.

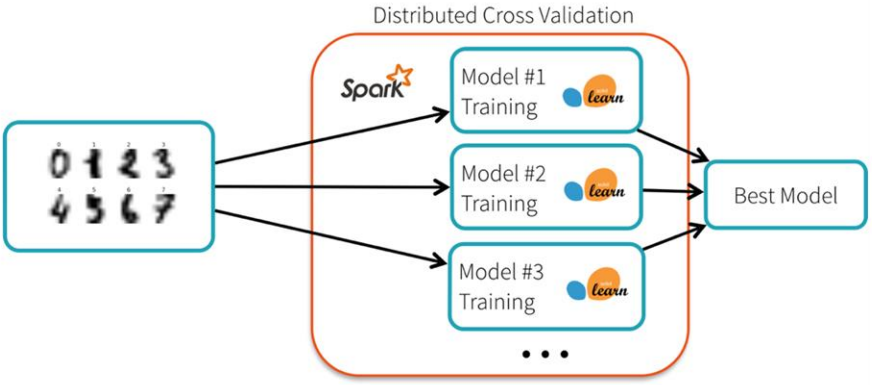
Module# 75: SciKit Learn Python Library

Scikit Learn is a free software machine learning library for the Python programming language. It features various algorithms:

- Classification
- Regression
- Clustering Algorithms

Scikit Learn interoperate with Python Numerical & Scientific Libraries (NumPy & SciPy), it Support :

- Vector Machines
- Random Forests
- Gradient Boosting
- k-means
- DBSCAN



CS442- Introduction of Data Science

Module# 76: Theano Python Library

Theano is a Python library and optimizing compiler for manipulating and evaluating mathematical expressions, especially matrix-valued ones:

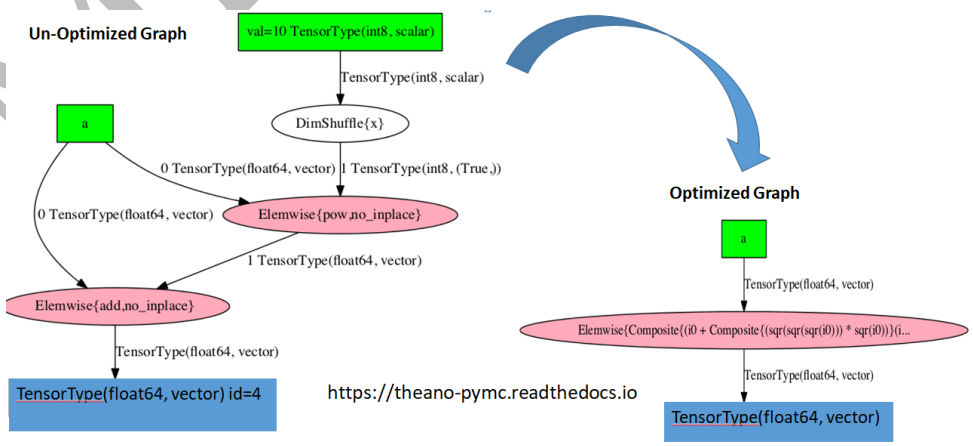
- Computations are expressed using a NumPy-esque syntax
- Compiled to run efficiently on either CPU or GPU architectures
- Extensive Unit Testing & Self Verification

Theano allows us to evaluate mathematical operations including multi-dimensional arrays so efficiently:

- Best to solve problems involving large amounts of data
- It take structures and convert them into very efficient code using other libraries (C Code)
- Handle computations required for large neural network algorithms

Theano represents symbolic mathematical computations as graphs composed of interconnected Apply, Variable and Op nodes:

- **Code**
- import theano.tensor as tt
- x = tt.dmatrix('x')
- y = tt.dmatrix('y')
- z = x + y



CS442- Introduction of Data Science

Module# 77: Keras Python Library

Keras is an open-source software library that provides a Python interface for artificial neural networks.

- Keras acts as an interface for the TensorFlow library
- Designed to enable fast experimentation with deep neural networks
- User Friendly & Extensible

Keras supports multiple backends, including:

- TensorFlow
- Microsoft Cognitive Toolkit
- Theano
- PlaidML
- Build-in DL Models

Keras Benefits:

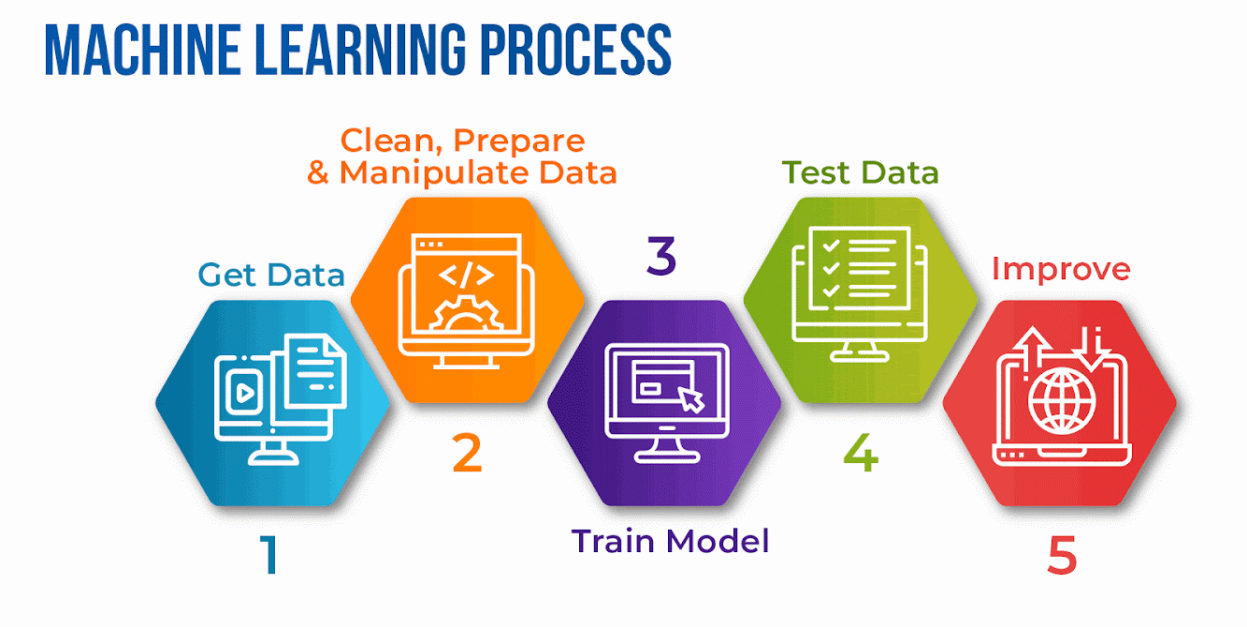
- Easy & Enjoyable
- Faster & Productive
- Strong Distributed Support
- Multiple Platform Models
- Large Ecosystem of Products

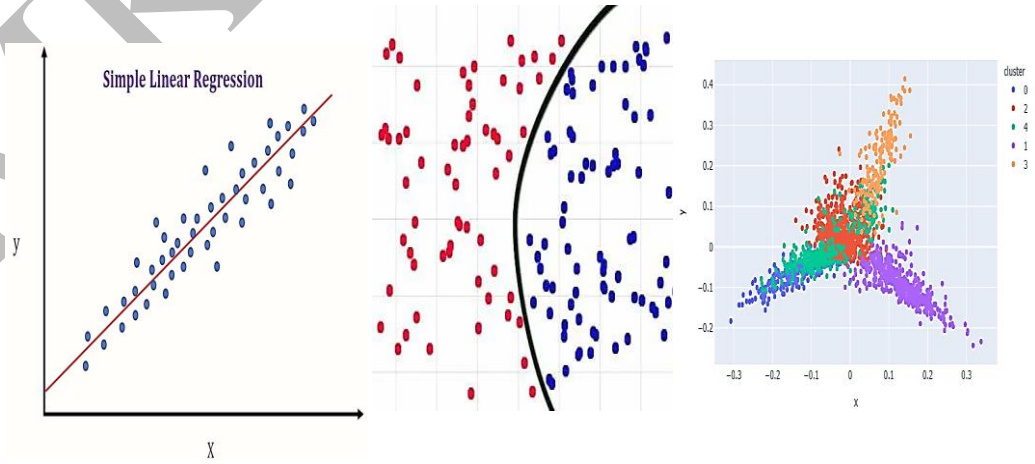
CS442- Introduction of Data Science

Module# 78: Machine Learning In Data Science

Machine Learning Algorithms are used in the Data Science Lifecycle:

- Machine Learning automates the process of Data Analysis
- Makes data-informed predictions in real-time
- A Data Model is built automatically and further trained to make real-time predictions





Regressio

Clssification

Clusterin

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science

CS442- Introduction of Data Science