Day 2

Data Engineering Pipelines
- Data Volume
- Data Velocity
- Data Variety
- Data Value
- Identifying Big Data Sources

Probability and Statistics
- Fundamental Concepts of Probability
- Probability and Inferential Statistics
- Random Variables
- Probability Distributions
- Expectations

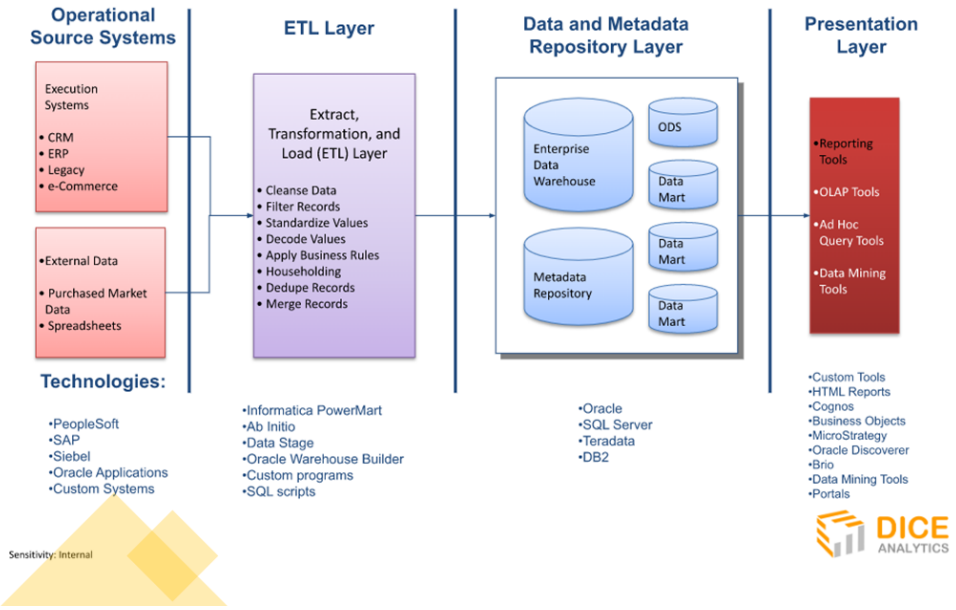
NumPy
- Understanding Data Types in Python
- Fixed-Type Arrays
- Arrays from Lists
- Arrays from Scratch

Data Engineering Pipelines:

Data Engineer => Data Science + Data Engineering
Skills => Prgmg skills, data storage (DB),System Implemetation, Database Management
Data Scientist => AI, ML , DL , Math , statistics, programing language , domain knowledge

Data Engineering => It will help to create a dataset for Data Modelling and data production and send this data to Data Science dept **(Data Scientist)**

Data Engineering Pipelines



Big Data: It is a combination of Structured (Relational DB -> SQL , Oracle) , unstructured (Non Relational , No SQL DB → Cassandra , Reddis) and semi structured data (Unstructured DB)collected by a company
Eg: FB
Eg: Pig, Hive , Hbase etc …

5V's:

- Velocity - How quick data generates & moves
- Volume - Refers to the size of data
- Value - Check whether the data is valuable or not

- Variety - It will analyse the data types and storage db like relational or non relational db
- Veracity - This defines the quality of data

Probability and Statistics:

Fundamental Concepts of Probability

- For solving the prediction problem -> Probability is used
- In DS:
  a. Data preprocessing
  b. model evaluation
  c. Visualisation of features (numpy , pandas)
  d. Dimensionality reduction
  e. Feature engineering
  f. Mean or Expectation value
    - Possible outcome of a random experiment repeated again and again for n times is called Expectation value
      Eg: Six face dice
      Possible outcome: {1,2,3,4,5,6} =⅙

      EV = 1(1/6)+2(1/6)+3(1/6)+4(1/6)+5(1/6)+6(1/6)
      = (1+2+3+4+5+6)/6 = 21/6 = 3.5
      EV of 2 dice = 3.5+3.5 = 7
      EV of n => 3.5 *n
      Assignment -> Find the EV of 1 dice of 4 faced =?
      (1+2+3+4)/4 =10/4 = 2.5

EV of 3 dice of 4 faced = 7.5

- Variance
  a. Find the mean of the given data set. Calculate the average of a given set of values
  b. Now subtract the mean from each value and square them
  c. Find the average of these squared values, that will result in variance
     Eg: 610,
     a.Mean = 1950 / 5 = 390
     b.(610-390)^2+(450-390)^2+(160-390)^2+(420-390)^2+ (310-390)^⅖

     =220^2 +60^2 +(-230)^2 +30^2 +(-80)^2/ 5

     =48400+3600+52900+900+6400 / 5
     = 112200/5
     = 224400

- Standard deviation
  1.Mean
  2.Diff of value with mean, square and sum
  3.Square root of o/p of step 2 / (sum-1)
  Eg: 4,2,5,8,6
  1.25/5 = 5
  2.(4-5)^2+(2-5)^2+(5-5)^2+(8-5)^2+(6-5)^2
  = (-1)^2+(-3)^2+(0)^2+(3)^2+(1)^2
  =1+9+0+9+1 =20

3. Square root of (20) / (5-1=4) = 20/4=5
Square root of 5 = 2.23

- Bayes Theorem
  Eg: 1 Image -> 4 boys , 6 girls
  2 Image ->4boys , 3 girls

  Find 1 person in Image 1 -> Girl?

  Image 1 => ½
  Image 2=> ½
  Prob of finding a girl in image 1 => 6/10
  Prob of finding a girl in image 2 => 3/7

  $$BT = \frac{½ * 6/10}{(½ * 6/10) + (½ * 3/7)} = 7/12$$

Probability and Inferential Statistics:

DS -> AI, ML , DL , Math , stat, Domain Knowledge, Prgmg kn

Eg: **Gender Classification**

Data(I/ p Images) -> Apply Algo(DL, ML , Math algo) -> Creating model(Brain) -> Test image (1 f, 3 m)
1 f ⇒ Accuracy ->80%
3 m=> Accuracy ->75%

Random Variables

1.It is a numerical data of random phenomenon
2.It is function of **real num**

Real Numbers:

1.Whole Num:(starts from 0…n)
2.Natural Num:(starts from 1…n)
3.Rational Num:(a/b)->3/6
4.Irrational number:square root of 3
5.Intergers:(-infinity …….+infinity)

Eg:
22/7 -> Rational number
3.14 -> Irrational number

Eg: Cricket :(1 over => 6 balls)
Prob of entire over => Range X=>{0,1,2,3,4,5,6}
Prob of Each delivery=> ½ (Caught, Not caught)
Total number of outcomes =>2^6 =64
No of outcome with no catch in a over=> 1/64
Random variable=> {0,1,2,3,4,5,6}

a.Discrete Random variable

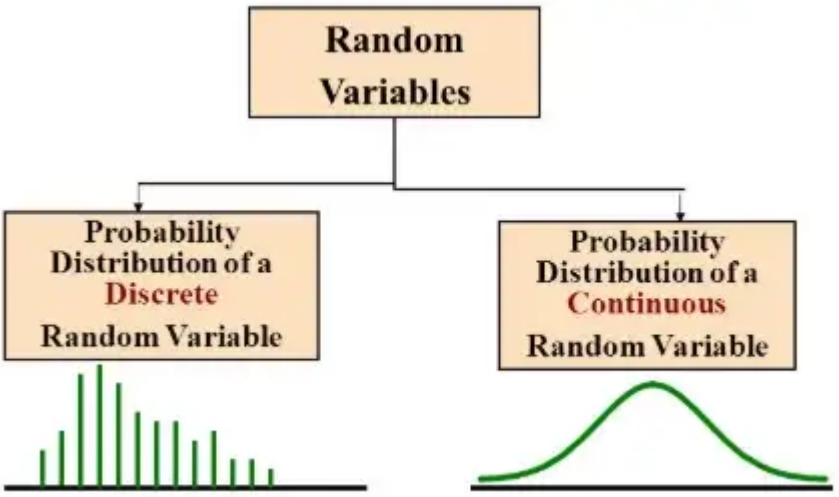
Countable value / Finite value of random variables

b.Continuous Random Variable

It is a random variable of infinite number of possible values

Eg: T20 Match
Person => 4 over
Find the bowling speed of a bowler in 4 overs?
4 overs = 24 balls

Bowling speed of 1st ball = 136.7km/h
Bowling speed of 2nd ball = 110.5km/h
Bowling speed of 3rd ball = 140.3km/h
{136.7, 110.5, 140.3, …….}



Distribution:

It is a visual representation of data to understand the o/p clearly

a.Discrete distribution

Numpy , matplotlib , plotly , scipy, Seaborn

4 types:

1.Uniform distribution ->flower shop (min , max)

2.Binomial distribution -> 2 possibility

Eg: Coin tossing

3.Poisson distribution(Time interval) -> Call center : how many calls a person can get?

Eg: 10.17,10.18,10.19, 10.20

C1=>(10.17 -10.18)

4.Geometric distribution-> points, lines, shapes , angles

b.Continuous distribution

Infinite num to be visualised in mat, plotly

Probability Distributions

Eg: Local election:

Parties-> P1, P2, P3, P4
Find the prob of winning team by non voters
100 people in district ->P1
10 -> Opposing P1
100 - 10 =90 -> winning score of P1

Expectations Value

NumPy

- Understanding Data Types in Python
- Fixed-Type Arrays
- Arrays from Lists
- Arrays from Scratch