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

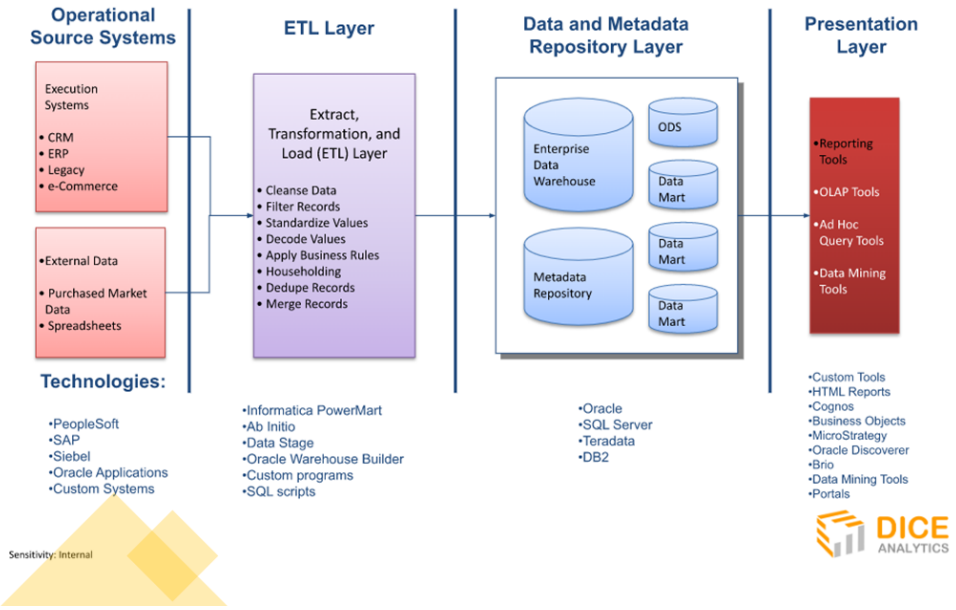
NumPy
- Understanding Data Types in Python
- Fixed-Type Arrays
- Arrays from Lists
- Arrays from Scratch

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
Random Variables
Probability Distributions
Expectations

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