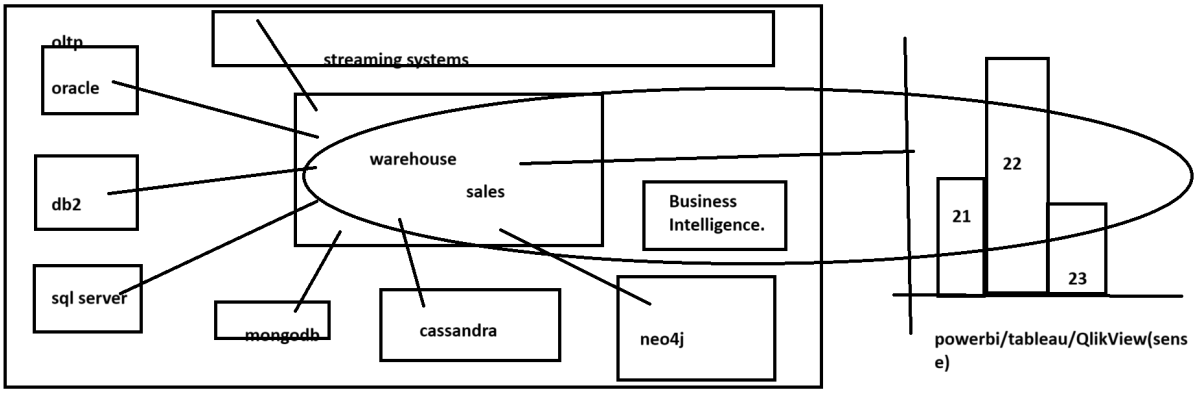
data engineer vs data analyst vs data scientist ?
---

data engineer -->
- ---> data extractions from different sources.
  - -->file systems (onpremise/ cloud)
  - --> rdbms (onprem/cloud)
  - --> nosql (mongodb/cassandra/neo4j...)
  - ---> data warehouse (onprem/cloud)
    - ---> ex onprem datawarehouse -->
      - teradata, netezza,vertica
    - --> ex for cloud datawarehouse
      - azure synapse analytics -- microsfort
      - snowflake
      - bigquery -- gcp
      - redshift -- aws
- --> structured
- --> unstructured
- ---> semi structured
  - --> unformatted ---> formatted(structured).

- ---> handling streaming data
  - ---> continous flow of data.
    - ex: IoT enabled censors
      Live Cams
      application logs, web logs, db logs etc.

  - ---> capturing streaming and perform real time analytics.

data engineer -->
  provisions data to different departments.
- --> Business Intelligence
- --> data analytics ()
- --> data scientist
- --> AI team.



#----------

below questions can answer you limitations of a Business Intelligence team.

question 1:

---

explain me sales report of recent 3 years (2021 to 2023)

---> bi can do ? (yes)

question 2:

---

compare the growth rate and growth status of recent 3 years sales report.

---> bi can do ? (yes)

question 3:

---

why in 2023 sales volume decreased?
what are top 10 reasons for that ?

--> BI can not.
--> solution by Data analyst.
--> By applying some mathematical/statistical models,

he will find out influence power of each parameter(reasons)
on target variable.
finally selects top 10 reasons.

--> predictions and forecasting.

data analyst limitations:

----> can not predict target variables,

if the data has complex patterns.

case 1

Image analytics.

Image ?

3 dimensional pixel matrix--> 100X100
r --> 100X100 --> 10000
g -->100X100
b -->100X100

10000X3 ---> 30000 pixels.

what is each pixel --> is 0 to 255 number

1000 male pics
1000 female pics

What data analyst with statistics can not predict ?

statistitian --> predict gender (predict gender based on image accurately ?)
--> to predict mood of a person
->moods like happy,sad,angry ?
is it possible by regular statistical techniques ?

---

case 2:
---
task : input is voice , based on voice of human , predict gender.

voice ---> m/f

task: input is voice, based on voice of any being, predict human or animal or bird
voice ---> human/animal/bird

case 3:
example: if you do some search on google,
input is few text of characters and output by google is sequence of set of words.
input sequence of text ---> next sequence of words

Above are limitations of Data analyst with statistical techniques.

#---

if more complex patterns involed in input data:
  --> statistical predictical models will give less accuracy.
          ex: 60% -- 70%

complex data ?

    High commonality and less distincted features.
   --> this type data is called "Complex".

above complex patterns can not be dealt with statistical techniques(models)

solution ---> is by "Data Scientist"
    --> He will be using "Statistics"
                          "Machine Learning"
                          " Neural networks"
                          " Deep Learning"

    by using above 4 desciplines, He will be doing predictions by building trained Models.

then how to decide a best model descipline ?

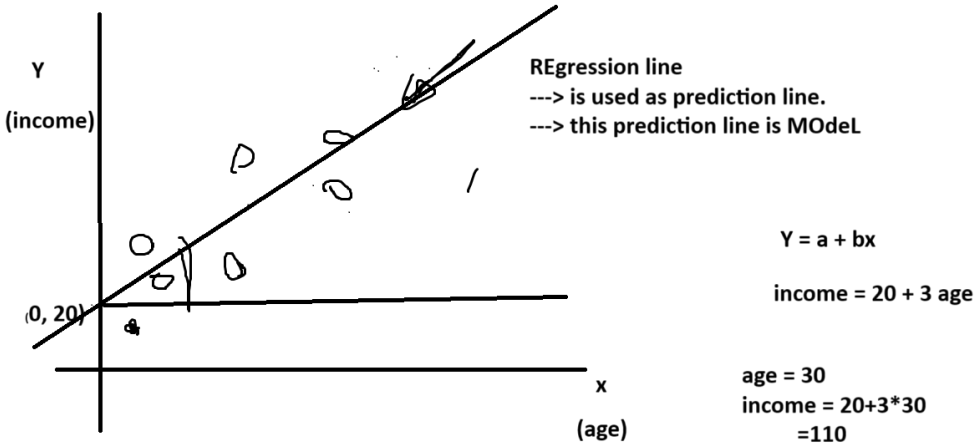
you build a model
  with
          1. stat ---> accuracy : 50%
          2. ML ---> accuracy : 70%
          3. NN ---> accuracy : 90%
          4. DL ---> accuracy : 97%

for example, Deep Learning (DL) has more accurcy (97%).

so we should deploy and use DL model (it has given high accuracy)

Model ?
----
what is model ?

model is a mathematical equation (the equation should give prediction)
--> model is mathematical predictive equation.



ex:
y=mx+c is a math equation , which produces straight line.
but every straight line is not a model.

but there is one straight line, which is optimized closest distance to all data points. that line is called regression line.

this regression line gives prediction based on input variable value.

this is called "model".

model is a Mathematical predictive equation.

y = mx + c --> math equation ---> not a model
x --> input variable
y ---> output variable
m ---> slope
c ---> constant

y = 2 x + 3 ---> straight line.
x=1 --> y = 5 ---> (1, 5)
x=20 --> y = 43 ---> (20, 43)
slope --> 2
constant --> 3

every math equation is not a model.

in statistics,
for prediction , we have "Regression Equation"

y = a + b x ---> regression line (math equation) ---> model

x ---> input variable
y --> target variable
a --> intercept
b ---> slope

ex: task
based on age predict income
input variable --> age
target variable --> income

y = a + bx

income = a + b. age

a -->intercept
b --> slope
(simply above two are parameters).
the parameters(a, b) are constructing relationship between input variable and output(target) variable.

y = b0 + b1. x1 + b2 . x2 + ... + bn. xn

x1,x2,x3...,xn ---> input variables (how many --> n)
y ---> target variable
b0 ---> intercept
b1 ---> slope of x1
b2 ---> slope of x2
:
:
bn ---> slope of xn

#---

#---
ml --->
is an approach to train models(math predictive equation)

differece between stat and ml ?

-- in next class we will discuss .