Q What is Data?

The Quantities, characters or symbols on which Operations are performed by a Computer which may be stored & transmitted in the form of electrical signals & recorded on magnetic, optical or mechanical recording media.

Q What is Big data?

Big data is a collection of data that is huge in volume yet growing exponentially with time. It is a data with so large size & complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size

Big data analytics

It largely involves collecting data from different sources, manage it in a way that it becomes available to be consumed by analysts & finally deliver data products useful to the Organization business.

Data Analyst

- The process of the extraction of information from a given Pool of data is called Data analytics.. A data analyst is a person who engaged in this form of analysis.
- A data Analyst extracts the information through several methodologies like data Cleaning, data Conversion & data modeling.
- data analytics allows the industries to process fast Queries to produce actionable results that are needed in a short duration of time.
- This restricts data analytics to a more short-term growth of the industry where Quick action is required.

Data Analyst Role

- Gathering data, using database query languages to retrieve & manipulate information.

- Perform data filtering, cleaning & early-stage transformation.
- Communicating results with the team using data visualization.
- Converting raw data into visual form.
- Data visualization is their key role.
- Work with the management team to understand business requirements.

Data Engineer :-

- A data engineer is a person who specializes in preparing data for analytical usage.
- Data Engineering also involves the development of platforms & architectures for data processing.
- Data engineers have to deal with Big data where they engage in numerous operations like data cleaning, management, transformation, data duplication etc.
- A data engineer is more experienced with core Programming concepts & Algorithms.

Eg: Developing a cloud infrastructure to facilitate real time analysis of data requires various development principles. Therefore, building an interface APIs is one of the job responsibilities of a data engineer.

Tools Used By data Engineers

① Hadoop ② Spark

Data Engineer Responsibilities

- Development Construction & maintenance of data architectures
- Conducting testing on large scale data platforms.
- Handling error logs & building robust data pipelines.
- Ability to handle raw & unstructured data.
- Ensure & support the data architecture utilized by data scientist & analysts.

- Development of data process for data modeling, mining & data production.

Data Scientist

- Data scientists are analytical experts who utilizes their skills in both Technology & social Science to find trend & manage data. They use industry knowledge, Contextual understanding, skepticism of existing assumptions — to uncover solutions to business challenges.
- A data Scientist's work typically involves making sens of messy unstructured data, from sources such as smart devices, social media feeds, & emails that don't neatly fit into a database.
- Data scientists are big data wranglers, gathering & analyzing large sets of structured & unstructured data.

Data scientist Responsibilities :-

- Performing data preprocessing that includes data transformation as well as data cleaning
- Using various machine learning tools to forecast & classify patterns in the data.
- Increasing the performance & accuracy of ML algorithms through fine-tuning & further performance optimization.

Applications of Big data in 10 Industry verticals.

(1) Banking And securities challenges :-

(i) early warning for securities fraud & Trade visibility
(ii) Card fraud detection & audit trails.
(iii) enterprise Credit risk reporting.
(iv) Customer data transformation & analytics.

(2) Communication, Media & entertainment :-

Challenges :-

(1) Collecting, analyzing & utilizing consumer insights.

(2) Leveraging mobile & social media contents
(3) Understanding patterns of real time media contents usage.

(3) Healthcare providers:
Challenges
(1) Rising medical costs.
(2) Unavailability / inadequate / unusable data.

(4) Education
Challenges:
(1) Incorporating data from varied sources.
(2) Untrained staff & Institutions about Big data.
(3) Issues of privacy & data protection.

(5) Manufacturing & Natural Resources
Challenges:
(1) Increase in the volume, complexity & velocity of data due to rising demands of natural resources.
(2) large volumes of uncapped data from the manufacturing industry.
(3) underutilization of data prevents improved quality, energy efficiency reliability & better profit margins.

(6) Government
Challenges
(1) Integration
(2) Interoperability of big data.

(7) Insurance
Challenges:
(1) lack of personalized services, pricing, targeted services to new market segments.
(2) Underutilization of data gathered by loss adjusters.

(5) targets for better insight.

Conventional System:
System which consist of 1 or more size each having either manual operated or automatic detection devices or combination of both. Big data is a huge amount of data which is beyond the processing capacity of conventional data base system to manage & analyse the data in specific time interval.

Difference b/w Conventional & Intelligent Computing
- The Conventional Computing functions logically with the sets of rules & calculation while intelligent computing can function via images & concepts.
- Conventional Computing is often unable to manage the variability of data obtained in real world. whereas intelligent computing is well suited to situation that have no clear algorithm solution and are unable to manage noisy imprecise data, this allows them to excel in those area that Conventional Computing often finds difficult.

Diff B/w Big data & Conventional data

Big data	Conventional data
(1) huge data set	(1) Data set size is in control
(2) Big data => unstructured data such as video, text, audio & images	(11) Normally structured data which has a fixed size & specific format.
(3) Hard to perform queries & analysis	(111) Relatively easy to perform query & analyse.
(4) need tools such as hadoop hive, hbase, pig, sqoop.	(4) tools such as sql & excel alone may be sufficient.
(5) use for reporting basic analysis & text mining Advance analytic is only in starting	(5) used for reporting & basic analytics modeling

(5) targets for better insight.

Conventional System:
System which consist of 1 or more size each having either manual operated or automatic detection devices or combination of both. Big data is a huge amount of data which is beyond the processing capacity of conventional data base system to manage & analyse the data in specific time interval.

Difference b/w Conventional & Intelligent Computing
- The Conventional Computing functions logically with the sets of rules & calculation while intelligent computing can function via images & concepts.
- Conventional Computing is often unable to manage the variability of data obtained in real world. whereas intelligent computing is well suited to situation that have no clear algorithm solution and are unable to manage noisy imprecise data, this allows them to excel in those area that Conventional Computing often finds difficult.

Diff B/w Big data & Conventional data

Big data	Conventional data
(1) huge data set	(1) Data set size is in control
(2) Big data => unstructured data such as video, text, audio & images	(11) Normally structured data which has a fixed size & specific format.
(3) Hard to perform queries & analysis	(111) Relatively easy to perform query & analyse.
(4) need tools such as hadoop hive, hbase, pig, sqoop.	(4) tools such as sql & excel alone may be sufficient.
(5) use for reporting basic analysis & text mining Advance analytic is only in starting	(5) used for reporting & basic analytics modeling

life cycle of data analytics unit (1)

The data analytic lifecycle is designed for Big data problems & data science projects. The cycle is iterative to represent real project.

(1) Discovery
- The data science team learn & investigate the problem.
- Develop context & understanding
- Come to know about data sources needed & available for the project
- the team formulates initial hypothesis that can be later tested with data.

(2) Data preparation
- steps to explore, preprocess & condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, & transform to get data into the sandbox

Page No.:
Date

- Data preparation tasks are likely to be performed multiple times & not in predefined order.
- several tools commonly used for this phase are - Hadoop, Alpine miner, Open Refine etc..

(3) Model planning

- In this phase data science team develop data sets for training & testing & production purposes.
- Teams build & execute models based on the work done in model planning phase

(4) Model Building :-

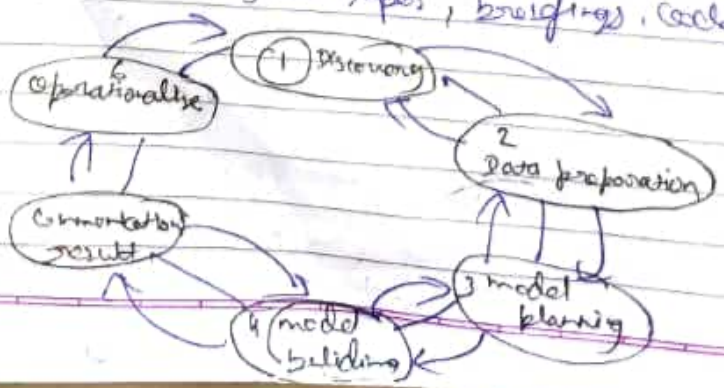
- Teams develops datasets for testing, training & production purposes
- Team also considers whether its existing tools will suffice for running the models or they need more robust environment for executing models.

(5) Communication Results :-

- After executing model team need to compare outcomes of modeling to criteria established for success & failure.
- Teams considers how best to articulate findings and outcomes to various team member & stakeholders, taking into account caveats, assumptions.
- Teams should identify key findings, quantify business value and develop narrative to summarize & convey findings to stakeholders.

(6) Operationalize :-

- The team communicate benefits of projects more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale & help & make adjustments before full deployment
- The team delivers final reports, briefings, code.



UNIT I

Types of data analysis

Analytics is the discovery, communication of meaningful pattern in data. Especially valuable in area rich with recorded info, analytics relies on the simultaneous application of statistics, computer programming & operation research to quantify performance.

Types
1. Predictive
2. Descriptive
3. Prescriptive
4. Diagnostic analytics

(1) Predictive Analytics :-

Predictive analytics turn the data into valuable actionable info. Predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning, data mining & game theory that analyze current & historical facts to make predictions about a future event. Techniques that are used for predictive analytics are,

- Linear regression
- Time series analysis & forecasting
- Data mining

Page No.:
Date:

Descriptive Analytics

Descriptive analytics looks at data & analyze past event for insight as to how to approach future events. It looks at the past performance & understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales marketing, operations, & finance uses this type of analysis. The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups unlike a predictive model that focuses on predicting the behaviour of a single customer, Descriptive analytics identifies many different relationship between customer & products.

Prescriptive Analytics

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction & then suggests a decision option to take advantage of the prediction. Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions benefits from the predictions & showing the decision maker the implication of each

decision option prescriptive analytics not only anticipate what will happen but prescriptive analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk & illustrate the implication of each decision option.
Eg - Prescriptive Analytics can benefit health care strategy planning by using analytics to leverage operational & usage data combined with data of external factors such as economic data, population demography, etc.

Diagnostic Analytics: In this analysis which generally, used historical data over other data to answ any que or for this solution of any problem we try to find any dependency & pattern in the historical data of a particular problem.
Eg: Companies go for this analysis because it gives a great insight into a problem & they also keep detailed information about their disposals otherwise data collection may turn out individual for very problem & it will be very time consuming.
Common techniques used for Diagnostic Analytics are:-
1) Data Discovery
2) Data Mining
3) Co-relations

Page No.
Date:

5 V's of Big data

Big data is a collection of data from many different sources & is often described by 5 characteristics: volume, value, variety, velocity & veracity.

- volume: the size & amounts of big data that companies manage & analyze
- value :- the most important V from the perspective of the business, the value of big data usually comes from insight discovery & pattern recognition that lead to more effective operations, stronger customer relationships & other clear & quantifiable business benefits.
- variety: the diversity & range of different data types including unstructured data, semi-structured data & raw data.
- velocity :- the speed at which companies receive, store & manage data, Eg the specific no. of social media posts or search queries received with a day, hour or other unit of time.
- Veracity = the truth or accuracy of data & the assets which often determine executive-level confidence.

Data Analytics

Date.
Page No.

Statics

Statics is a disiplare dealing with collecting analy-sing and interpreting finally presenting that numerical data. It is used in every aspects of data science, that means it is used to analyse, transformed & cleared data.

- Population is the entire data set which is used for a particular test
- sample is a portion of population which is used for testing.

Types of Statics

① descriptive ② Inferential

Descriptive Statics

- It helps us to describe the data
- It enables us to underline the characteristics of data
- It doesnot predict anything if it does not make any assumption

Eg mean, mode and median.

weight (Imbalance)

Date.
Page No.

Inferential Statics

- It helps us to draw conclusion i.e.
- It helps to predict or make assumptions.

Eg Histogram, Bar Chart

Limitation of Descriptive Statics.

- They are useful but they hide the important information about the data set. So we use inferential Statics.

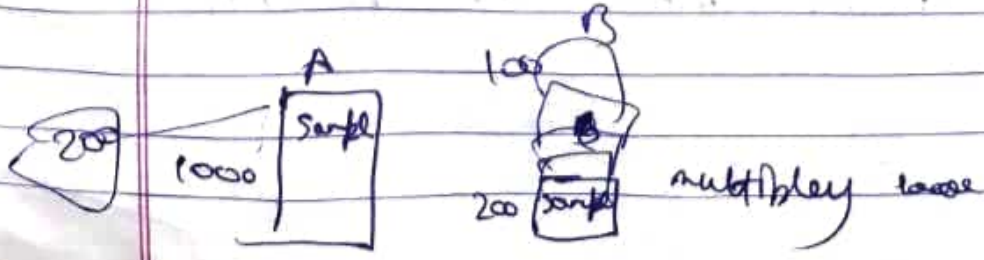
Probability = no of events / total no of outcomes.

Eg = Rolling of dice

$P = \frac{1}{6}$

Bias = tendency of a statistical or predictive model to over or underestimate a parameter

(1) Over sampling
(2) Under sampling



Date.
Page No.

Over sampling

we use to apply ML techniques to model the data and make prediction when the sampled is biased then we perform 2 types of sampling

(1) under sampling
(2) over sampling

(1) under sampling :

we select some data from the majority class as the same no of the minority class.

Over Sampling

multiply the minority class such that it has the count as the majority class.

what is sampling?

It is the active process of gathering observations with the intent of estimating a population.

Resampling : It is the methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter.

Date.
Page No.

Sc Sampling error = Calculated o/p
- Observed o/p

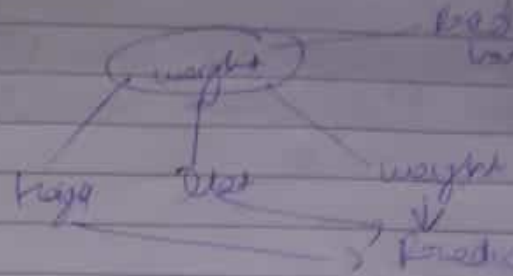
⇒ we can perform sampling by 2 method.
(1) Bootstrap method
(2) K-field cross validation

Bootstrap
Samples are drawn from the data set with replacement allowing the same sample to appear more than once in the sample

K-field cross validation
In K-field cross validation data set is partitioned into K-groups where each group is given the opportunity of being used as a sample.

Applications of Statics
(1) Business Analytics
(2) artificial Intelligence
(3) Financial Analysis
(4) Fraud detection
(5) Machine learning
(6) Share Market
(7) pharmaceutical section

Regression



Regression

"It is a method to mathematically formulate relationship b/w variable i.e. results in what manner the predicted variable is affected by the predictor.

Linear Regression

$y = bx + c$

Non linear regression → $y = b_1^2 x_1 \; b_1^2 x_2$

Logistic Regression → It is binary (2 values) Eg: Happy / Sad, True / false

Time series Regression → Based on Time

Regression Types

1. Linear Regression →

It assumes that there is a linear relationship b/w predictors (x) & predicted variable (y)

$y = b_1 x + c$

Multiple linear regression

It assumes that a predicted variable is based on more than 1 independent variable

eg: $b_0 + b_1 x_1 + b_2 x_2 + \dots + \epsilon$

Non linear regression

It assumes that relationship b/w x & y is having a polynomial function

Logistic Regression

It is useful when a target variable is binomial (1 or 0) (accept or reject)

It estimates probability of occurring event

Eg: Happy or Sad, Rain/Fall.

It works if it has value 0 or 1.

Time Series regression

It is used to forecast behaviour of a variable based on historical time

Eg: unemployment rate

$Y_t = \gamma_1 + \gamma_2 t + \gamma_3 t$

Y_{t-1} : It has value of Y

Page No.
Date

(1) Calculate prediction error & Root mean square error for the data

Predicted value [y]	Actual value [Y]	$(Y-y)^2$
14	12	4
15	15	0
18	20	4
19	16	9
25	20	25
18	19	1
12	16	16
12	20	64
15	16	1
21	15	36
		160

$$RMSE = \sqrt{\Sigma (Y-y)^2} = \sqrt{160} = 4\sqrt{10}$$
$$\simeq 4 \times 3.14$$
$$\simeq 12.56$$

for logistic regression

Misclassification rate = $\frac{\text{Incorrect prediction}}{\text{total prediction}}$

Predicted	Actual	Incorrect
1	0	✓
1	1	
1	1	
0	1	✓
0	0	
1	0	✓
0	0	

∴ Misclassification rate $= \frac{3}{7} \times 100 = 42.8\%$

Multivariate Analysis

It involves multiple dependent variables resulting in one outcome.

Uses :-

It is widely used in many industries like healthcare. Recently on COVID a team have more than 5 lakh covid patients till end of July 2020. This analysis was made for multiple variable like

- Govt decision
- Public behaviour
- Corruption
- Public transport
- healthcare services
- Overall immunity of Country

Advantages

- Therefore it considers more than 1 factor of independent variables. Hence the conclusion drawn is more accurate.
- Conclusion are more realistic.

Disadvantages

It requires complex computation to arrive at a statistical conclusion. Many observations for a large no. of variables leads back. Affected if data is lacked.

Bayesian modelling => It is a statistical approach that combines the different pieces of probability to solve statistical problems.

Event => It refers to the forecasted likely to happen

Posterior => The probability of occurrence of event on basis of prior knowledge

Bayes Rule =>

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} \quad \text{--- (1)} \qquad P(B) \neq 0$$

Conditional probability $P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$ --- (2)

Using (1) & (2)

likelihood

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$ --- Bayes

P(A) — prior

P(B) — evidence

Posterior

likelihood =>

Probability of event B being True given that event A is true.

It refers to the probability of observing what we did given that our prior are true

Eg => A situation where bayesian analysis is used is the spam filter in your mail server. A message is scanned used for the occurrence of keywords which have a likely role in classifying the message as spam.

Q. Let S → message is spam
C → Subject line contains the sentence "Check this out"

Compute conditional probability P(S|C) where 40% of emails are spam, 1% of spam emails have "Check this out" in the subject line, 0.4% of the non-spam emails have this sentence in the subject line.

Sol. $P(S) = \frac{40}{100} = 0.4$

$P(C/S) = \frac{1}{100} = 0.01$

$PC = \frac{0.6 \times 0.4}{100} = 0.0024$

$P(S/C) = \frac{P(C/S) \cdot P(S)}{P(C)}$

$PS/C = \frac{0.001 \times 0.4}{0.0024} = \frac{0.001 \times 0.4}{0.0024}$

$\frac{40 \; 10}{24 \; 6} = \frac{10}{6} = 1.67$

Q. Find out conditional probability where
B → ...
A → ...

$P(A) = \frac{1}{2}$ $P(B) = \frac{1}{4}$

$P\left(\frac{B}{A}\right) \rightarrow \frac{1}{2}$

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{4}} = \frac{\frac{1}{4}}{\frac{1}{4}} = 1$$

Major advantage of Bayesian statistics

It takes the prior knowledge into consideration while calculating probability by applying Bayes rule.