File Handling -

The file is use to store the inforamtion

Such as : -

* TextFile
* Csv File
* Excel File
* JASON File
* Binary File
* XML File
* Video File
* Image File
* Audio File

getwd() -> This command is use to display the current working directory.

Ex –

Setwd() -> This command is use to set the directory to store the file.

Ex -

File.create -> This command is use to create the file in the current working directory.

TRUE - > If the output is true the file is successfully created in the current working directory.

Read.table() -> This command is use to read the content of the text file.

View() -> this command is use to display the data.

Difference between display and print

Print function display the content where cmd prompt or console.

View command function display the contect the separate file or script tab.

Create the csv file Reva.csv file

Doller Operator $ -> this opertaor is use to access the one column of the csv file.

Syntax – object\_name$column\_name

sub set() -> This command is use to perform many queries on csv file.

Get the person details having the max salary.

Get the person details having min salary.

Get all the details of people in working IT department.

Get all the details of people in working Finance department.

Get all the details of people in working of IT department and salary is more than 600.

On or after 2014

Get all the people who are working for Finance department who are join on or after.

Read the Excel File

Install xlsx Package

Cmd -> install.packages(“xlsx”)

Verify the package is install

Any(grepl(“xlsx”,install.packages()))

Read the File from worksheet

data = read.xlsx(“input.xlsx”,sheetIndex = 1)

print(data)

Binary File

A binary file is a file that contains infromation stored only in from of bits and bytes(0’s and 1’s).

They are not a human readable as a bytes in it translate to character and symbols which conatin many other non-printable charcters.

Syntax – 2 Function write and Read.

1. WriteBin(object,con)

2. readBin(con,what,n)

* Con is connection object to read or write the binary file.
* Object is the binary file which to be written.
* What is the mode like character , integer, etc re

Ex –

“cyl”, “am” and “gear”

Write.table(mtcars, file = “mtcars.csv”, row.names = FALSE, na = “ ”, col.names = TRUE, sep = “,”)

new.mtcars = read.table(“mtcars.csv”, sep = “,”, header = TRUE, row = 5)

write.filename = file(“/web/com/binmtcars.bat”, “wb”)

writeBin(colnames(new.mtcars), write.filename)

writeBin(c(new.mtcars$cy,new.mtcars$am,new.mtcars$gear),write.filename)

close(filename)

Read in the Binary File –

The Binary file cr

read.filename = file(“/web/com/binmtcars.bat”, “rb”)

column.names = readBin(read.filename, charcater(), n = 3)

print(bindata)

cyldata = bindata[4:8]

print(cyldata)

amdata = bindata[9:13]

print(amdata)

geardata = bindata[14:18]

print(geardata)

finaldata = cbind(cyldata, amdata, geardata)

colnames(finaldata) = column.names

print(finaldata)

**Databases**

RMySQL Package

install.packages(“RMySQL”)

mysqlconnection dbConnect(MySQL(), user = “system”, password = “alex” dbname = “saki”, host = ‘localhost’)

dbListTables(mysqlconnection)

Query in the Tables –

result = dbSendQuery(mysqlconnection, “select \* from actor”)

result = dbSendQuery(mysqlconnection, “select name from actor)

data.frame = fetch(result, n = 5)

print(data.frame)

get all the details of the actor of age >= 30

result = dbSendQuery(mysqlconnection, “select \* from actor where age >= 30”)

data.frame = fetch(result, age >= 30)

print(data.frame)

dbSendQuery(mysqlconnection, “update mtcars set disp = 168.5 where hp = 110”)

dbSendQuery(mysqlconnection, “insert into mtcars(row\_names, mpg,cyl,,disp,,drat,wt,qsec,vs,am,gea) values(“New Mazda RX4”,120,20,366)

dbwriteTable() – to create a table

mysqlconnection dbConnect(MySQL(), user = “system”, password = “alex” dbname = “saki”, host = ‘localhost’)

dbwriteTable(mysqlconnection, “mtcars”[, ], overwrite = TRUE)

dbSendQuery(mysqlconnection, “drop table if exist mtcars”)

Q - create a database student and table name student information with the attributes -

SRN

Name

Age

Subject\_Marks – Sub1,Sub2,Sub3

Total

Percentage

And write R statement in the above table.

1.Stablish the connection to the mysqlserver by using db cmd.

mysqlconnection dbConnect(MySQL(), user = “system”, password = “alex” dbname = “Student”, host = ‘localhost’)

2. result = dbSendQuery(mysqlconnection, “create table “Student”, [SRN varchar(10), Name varchar(20), integer(15),Sub1 integer(10),sub2 integer(10), integer(10)]”)

3. dbSendQuery(mysqlconnection, “insert into Student(SRN,Name,Age,sub1,sub2,sub3) values (“R179”,”Prince”,22,50,60,80)

dbSendQuery(mysqlconnection, “insert into

Student(SRN,Name,Age,sub1,sub2,sub3) values (“R180”,”Alex”,21,80,70,90)

Student(SRN,Name,Age,sub1,sub2,sub3) values (“R170”,”Rahul”,22,50,60,80)

dbSendQuery(mysqlconnection, “insert into

Student(SRN,Name,Age,sub1,sub2,sub3) values (“R160”,”Tushar”,21,80,70,90)

Student(SRN,Name,Age,sub1,sub2,sub3) values (“R129”,”MD. Maanish”,22,50,60,80)

dbSendQuery(mysqlconnection, “insert into

Student(SRN,Name,Age,sub1,sub2,sub3) values (“R130”,”PK”,21,80,70,90)

SRN total and of all

dbSendQuery(mysqlconnection, “select SRN, Name, Total from Student )

* display SRN, Name of all the student which perctentage more than 60

dbSendQuery(mysqlconnection, “select SRN, Name from Student where percentage > 60”)

* select first 3 rows and convert into the data frame.

result = dbSendQuery(mysqlconnection, “select \* from Student”)

data.frame = fetch(result, n = 3)

print(data.frame)

JSON File – JavaScript Object notation File

install rjson Packages

install.packages(“rjson”)

library(“rjson”) -> import the package

result = fromJSON(file = “input.json”)

print(result)

jsonf = as.data.frame(result)

print(jsonf)

XML File – Extensible Markup Language

install.packge(“XML”)

library(“XML”)

library(“methods”)

result = xmlParse(file = “input.xml”)

print(result)

rootnode = xmlRoot(result)

rootsize = xmlsize(rootnode)

print(rootsize)

UNIT – 2

Business Intelligents –

1. Business Intelligent (BI) market is a set of methodologies, processes, architecures and technologies that leverage the output of the infromation management processes for analysis, reporting, performance management , information delivery.
2. Business intelligence helps find answers to questions you know.
3. Big data helps you find the the questions you don’t know you want to ask.
4. Using AI tools, business can monitor the growing trends in the markets and address business problems as well as client quries.
5. BI is designed to handle static data and highly structure data.
6. BI store the data mostly in data warehouse.
7. In business intelligent, past data is analyzed to understand the current trends of the business.

**Business applications -> data warehouse -> analytics -> dashboard**

Data science encomposesses preparing for the analysis

* Including cleaning.
* Aggregrating
* Manipulating the data to perform advanced data analysis.

Data science is an interdisplinary field that

* Uses scientifices methods.
* Processess.
* Algorithms and system to extract knowledge and form noisy structure and unstructure data and apply knowledge and actionaable insights.

Big data is a techniques to

* Collect maintains
* Process the huge infromation
* It is about the collection, processing, analyzing and utilizations of the data of various operations.

Wisdom

|

Knowledge

|

Infromation

|

Data

||

Events

Q - What is business intelligent. And difference b/w business intelligent and data science.

|  |  |  |
| --- | --- | --- |
| Factor | BI | Data Science |
| Concept | Deals with data analysis on the business platform | Consists of several data operations in various domains |
| Scope | BI analyzes past data / Historical data | Past data is analyze for future predications. |
| Data | Handling static and stucture data | Both structure and unstructure data that is also dynamic |
| Data storage | Data stored mostly in data-warehouses | Data utilized is distributed in real time clusters |
| Procedure | BI helps companies to solve questions | Questions are both curated and solved by data scientists |
| Tools | MS Excel, SAS BI, Sisence, Microstratagy |  |
|  |  |  |

Q : - What are the roles of Data Scientists : -

* Data scientists deals with the data in order to assist companies in making proper decisions.
* A data scientists deals with both structure and unstructure data.
* The data scientists investigates organized data and analyzes it throughly to derive infromation from it using various statistical methodologies. Statical methods describe, visualize and hyphothesis infromation from the data.
* Advanced machine learning algorithms, the data scientists predicts the occourence of the events and take the data-driven methods.
* A data scientists deploys vast arrays of tools and practices to recognize redundant patterns within tha data. These tools range from SQL, Hadoop to Weka, R and Python.
* A data scientists usually act as consultants employed by companies where they participate in various decesion making process and creation of strategies.
* In others words data scientists use meaningful insights form data to assist companies in taking smarter business decesions.
* A data scientists should have a knowledge about the data source data preprocessing technique and data preprations technique.
* A data scientists should know how to integrate the multiple data source of different data fromat like : structure and unstructure in a single model.

Q : - What is Data analytics explain the diffference data analytics?

Types of Data Analytics : -

* Discriptive analytics.
* Diagnostics analytics.
* Predicative analytics.
* Prescritive analytics.

**Data Analytics Life Cycle : -**

* To address the distinct requirements for performaing analysis on big data, a step by step methodology.

Phase 1: Discovery. -> State problem and collect data.

Phase 2: Data Preprations. -> Data processing.

Phase 3: Model Planning. -> Explore / Transform data.

Phase 4: Model Building -> Build model.

Phase 5: Communication Results -> Communicate results / Publish insights.

Phase 6: Operationalize. -> Measure effectives / Apply live.

Correlation and Covariance these two are the relationship b/w object.

Covariance – E(xi – ~~x~~) (yi – ~~y~~) / n

Phase 1 : Discovery

* The data science team learn and investigate the problem.
* Develop context and understanding.
* Come to know about data sources needed and available to the project.
* The team formulates initial(null) hypothesis that can be later tested with data.

Phase 2 : Data Preprations

* Steps to explore, preprocess, and condition data prior to modeling and analysis.
* It requires the presence of analysis sandbox, the team execute, load and transform, to get data into the sandbox.
* Data preprations tasks are likely to be perform multiple time and not in predefined order.
* Several tools commonly used for this phase are – Hadoop, Alphine Miner, Open Refine, etc.

Phase 3 : Model Planning

* Team explore data to learn about the relationships between variable and subsequently, selects key variable and the suitable models.
* Data science team develop data sets for tranning, testing and production purposes.
* Team builds and executes models based on the works done model planning.

Phase 4 : Model Building

* Team develop datasets for testing and production purpose.
* Team also consider wheater its existing a

Phase 5 : Communication Result

* After executing models team need to compare outcomes of modeling to criteries establish for success and failure.
* Team consider how best to articulate

Phase 6 : Operationalize

* The team communication benefits of projects more

Distingush between analytics and reporting

|  |  |
| --- | --- |
| Analytics | Reporting |
| Analytics is the method of executing and analyzing summarized data to make business decisions | Reporting is an action that includes all the needed infromation and data and it put together in an organized way. |
| Questioning the data understanig it. Investigating it, and presenting it to the end users are all parts of analytics | Identifying businness events, gathering |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Google stored the data on cloud server - GFS (google file system)

Distributing computing & Parallal Processing System

Hadoop Eco System : - hadoop is a technology to store huge volume of differnet data size, different format data, at any point of time.

Define Hadoop EcoSystem ?

Expalin the different services with a neat diagram.

Hadoop Ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problem you can consider it as a suite which encomposses a number of services (investing, storing, and maintaing) inside it.

Hadoop components and tools of Hadoop EcoSystem –

HDFS – Hadoop Distributed File System. (Solving the Solution of Storage Problems)

* Sqoop – Structure Data (Store Data)
* Flume – Unstructure Data (Store Data)

YARN(Resource Management System) YARN – Yet Resource Negotiator (Resource Allocator) is main or core resource of Hadoop EcoSystem.

MapReduce – Data Processing using Programming. (Java is best design for MapReduce) - Data processing tools

Spark MLlib, Mahout– In Memory Data Processing. (data flow diagram) - ML

PIG, HIVE – Scripting language tools.

HBASE – NoSQL Database working.

Apache DRILL – Data Processing Services. – Analytical Tools (SQL-on-Hadoop)

Zookeeper & Ambari - to Manage(Monitor ~~or~~ Process the Clusters)

OOZIE – Scheduling the task.

SOLR & LUCENE – Searching and Indexing the task.

KAFKA & STORM – Streaming the task.

What is HDFS Explain with the suitable diagram.

* Core components or backbone of Hadoop EcoSystem.
* Store different types of large data sets like(structure data, unstructure data, semistructure data)
* Lebal of abstraction over the resources, whole HDFS as a single unit.
* HDFS store log file information stored data (meta data).
* HDFS has two components
* NameNode. (only the log file or meta data store)
* DataNode. (actual data store in DataNode)

**NameNode**

YARN – Yet Another Resource Negotiation.

* Consider YARN as the brain of your Hadoop EcoSystem.
* It performs all your processing activities by allocating and scheduling task.
* It has two components –
  + Resource Mangers and NameNode.
  + Node Mangers and DataNode.

Schedulars – Schedulers performs scheduling alogorithms and allocate the resources.

Applications Mangers – Accept the all Jobs and tasks

MapReduce –

It is the core components of processing in a Hadoop EcoSystem as it provides the logic of processing.

MapReduce is a framework which helps in writing applications that process the large data sets using distributed and parallal algorithms inside Hadoop environment.

Map() and Reduce() are two Function –

Map() Function is performs the actions like : filtering, grouping,

Reduce() Function is performs the

Input -> Split -> Map Phase -> Shuffle Sort -> Reduce Phase

Consider the following the data –

and 1

and 1

and or ok do ok

or 1

ok 2

ok 1

ok1

and, or, ok, do, ok, while, repeat, if, ok, do, and

While repeat if

ok do and

while 1

repeat 1

If 1

do 1

do 1

Apache PIG

It will support SQL commands

HIVE

HIVE + SQL = HQL

Hive command line and Jdbc and Odbc driver.

Apache Mahout

ML

Apache Spark

Real time data

Telecom data

Signal data

Sensor data

Apache HBASE

Apache DRILL

Apache Zookeeper

Apache OOZIE

Apache FLUME

Apache SQOOP

Linear Regression

Assume that there is only one independent variable x.

If the realtionship between x (independent varibale) and y (dependent variable or output varibale)

Y = a,b

~~X~~ = 1 / n Exi

~~Y~~ = 1 / n Eyi

Calculate the variance of X

Variance(x) = 1 / n -1 E(xi – ~~x~~~~i~~~~)~~~~2~~

Cov(x,y) = 1 / n – 1 E(xi - ~~x~~i)(yi – ~~y~~i)

b = cov(x,y) / var(x)

a = ~~y~~ - b~~x~~

WAP to Linear Program.

x = (10,2,3,6,12,7,8)

y = (20,12,15,16,7,4,10)

**Profile of data scientist**

Discription –

* Here a datascientist mentions the discriotions about the company name and innoviative data driven technology, n to n data analysis life cycle of the big data.
* The data scientist the processess a strong analytical ability to solve complex problems the data.
* As a data scientist at (data company name) you will drive real-world impact by applying innovative data-driven

Responsibilities –

* Work with large data sets on extraction, cleaning, explorations, analysing and presentation.
* Formulate data science solutioins of business problems collaboratting with business and product teams.
* Design and implement algorithms to solve as wide array of challenging problems with analytics and statical approches builts on high-volume and high-dimensional.
* Deploy, test, validate and maintains machine learning models in productions by collabrating with data engineers and machine learning engineers.
* Performs extract, transform, load operations from data sources for modeling purpose.
* Design perform and analysis A / B tests.

Experience –

* Bachler’s degree or higher, or equivalent practical experience.

Minimum qualifiacations –

* Solid knowledge of statical methods and machine learning fundamentals.
* Strong proficiency in python and R.
* Strong Proficiency in SQL.
* Hans-on experience with data analysis and classical machine learning tools.(ex- pandas,numpy,scikit-learn)
* Excellent written and verbal communications skills.
* Excellent stakeholder managements skills.
* Ability to tell stories with data and deliver data-driven insights to wide range of audience.

What will make you stand out –

* Experience in productions software engineering routines.
* Familiarity or experience with working on large data sets and distributed computing.
* Working knowledge of cloud-based solutions.

Difference between big data and cloud computing

|  |  |  |
| --- | --- | --- |
| Parameter of comparison | Big data | Cloud compiuting |
| Basic | It refers to the processing of massive amounts of data using a varity of methods to organise,store, examine, and maintains the data. | Utilizations of computer services such as storage server software network and analytics is what this refers to |
| Characteristics | The factors of volume diversity, velocity, authenticity and value are all taken into considersions | Adapatibility cost reductions idenpendence form devices and locations and simplicity of mainatainence multitendency grater productivity and security are some of benfits |
| Functions |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Business intelligence vs data science

Logistic Regression –

* Logistic Regression is one of the most popular algorithms, which comes under the supervised learning technique.
* It is predicting is predicting the catogerical depedent variables using a given set of independent variables.
* Logistic Regression predicts the output of a catogorical dependent variable. It gives the possibilities values lies between 0 and 1.
* It is instead of fitting a regression line, we fit an ‘S’ Shaped Logistic Function, which predict two maximum values (0 or 1).
* It is used for solving the classification problems. Linear Regression is used to solving the Regression Problems. It is similar to the Linear Regression expect that how they are used.
* Logistic Regression plot the value in form of S shape. It is also called Sigmoid Line. The value problablistic may be lies between 1 to 1.

Logistic Function – (Sigmoid Function)

* The sigmoid function is a mathematical function used to map the predicted values to probabilities.
* The value
* In the Logistic Regression, we use the concept the
* In Logistic Regression Y values lies between 0 and 1.
* Calculate the value y / 1 – y. if it is 0 – it is idle state and 1 – it is infinite state.