

NGM COLLEGE
PG DEPARTMENT OF COMPUTER SCIENCE
COURSE TITLE: BIG DATA ANALYTICS
COURSE CODE:18PCS208

K1 LEVEL QUESTIONS

UNIT-I

1. Hadoop is a framework that works with a variety of related tools. Common cohorts include:

- a) MapReduce, Hive and HBase
- b) MapReduce, MySQL and Google Apps
- c) MapReduce, Hummer and Iguana
- d) MapReduce, Heron and Trumpet

Answer:a

2. _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

- a) MapReduce
- b) Mahout
- c) Oozie
- d) All of the mentioned

Answer: a

3. Which of the following genres does Hadoop produce ?

- a) Distributed file system
- b) JAX-RS
- c) Java Message Service
- d) Relational Database Management System

Answer: a

4. _____ class adds HBase configuration files to its object.

- a) Configuration
- b) Collector
- c) Component
- d) None of the mentioned

Answer: a

5. _____ communicate with the client and handle data-related operations.

- a) Master Server
- b) Region Server
- c) Htable
- d) All of the mentioned

Answer: b

6. _____ is the main configuration file of HBase.

- a) hbase.xml
- b) hbase-site.xml
- c) hbase-site-conf.xml
- d) none of the mentioned

Answer: b

7. HBase uses the _____ File System to store its data.

- a) Hive
- b) Imphala
- c) Hadoop
- d) Scala

Answer: c

8. Which of the following class is provided by Aggregate package ?

- a) Map
- b) Reducer
- c) Reduce
- d) None of the mentioned

Answer: b

9. _____ is the architectural center of Hadoop that allows multiple data processing engines.

- a) YARN
- b) Hive
- c) Incubator
- d) Chuckwa

Answer: a

10. What is Big Data?

- Large volume of data
- Large velocity of data
- Large variety of data.

UNIT-II

11. YARN's dynamic allocation of cluster resources improves utilization over more static _____ rules used in early versions of Hadoop.

- a) Hive
- b) MapReduce
- c) Imphala
- d) All of the mentioned

Answer: b

12. The _____ is a framework-specific entity that negotiates resources from the ResourceManager

- a) NodeManager
- b) ResourceManager
- c) ApplicationMaster
- d) All of the mentioned

Answer: c

13. Apache Hadoop YARN stands for :

- a) Yet Another Reserve Negotiator
- b) Yet Another Resource Network
- c) Yet Another Resource Negotiator
- d) All of the mentioned

Answer: c

14. MapReduce has undergone a complete overhaul in hadoop :

- a) 0.21
- b) 0.23
- c) 0.24
- d) 0.26

Answer: b

15. HBase is a distributed _____ database built on top of the Hadoop file system.

- a) Column-oriented
- b) Row-oriented
- c) Tuple-oriented
- d) None of the mentioned

Answer: a

16. HBase is _____, defines only column families.

- a) Row Oriented
- b) Schema-less
- c) Fixed Schema
- d) All of the mentioned

Answer: b

17. Apache HBase is a non-relational database modeled after Google's _____

- a) BigTop
- b) Bigtable
- c) Scanner
- d) FoundationDB

Answer: b

18. Hadoop is a framework that works with a variety of related tools. Common cohorts include

- a. MapReduce, Hive And HBase
- b. MapReduce, MySQL and Google apps
- c. MapReduce, Hummer and Iguana
- d. MapReduce, Heron and Trumpet

Answer: a

19. How is Hadoop related to Big Data?

Hadoop is an open-source framework for storing, processing, and analyzing complex unstructured data sets for deriving insights and intelligence.

20. Define HDFS and YARN, and talk about their respective components?

The HDFS is Hadoop's default storage unit and is responsible for storing different types of data in a distributed environment.

HDFS has the following two components:

NameNode – This is the master node that has the metadata information for all the data blocks in the HDFS.

DataNode – These are the nodes that act as slave nodes and are responsible for storing the data.

The two main components of YARN are –

ResourceManager – Responsible for allocating resources to respective NodeManagers based on the needs.

NodeManager – Executes tasks on every DataNode.

UNIT-III

21. Facebook Tackles Big Data with _____ based on Hadoop

- a. Project Prism
- b. Prism
- c. ProjectData
- d. ProjectBid

Answer: a

22. _____ jobs are optimized for scalability but not latency.

- a. MapReduce
- b. Drill
- c. Hive
- d. Oozie

Answer: c

23. In HDFS the files cannot be

- a. read
- b. execute
- c. delete
- d. archived

Answer: b

24. Which of the following partitions the key space ?

- a) Partitioner
- b) Compactor
- c) Collector
- d) All of the mentioned

Answer: a

25. The files that are written by the _____ job are valid Avro files.

- a) Avro
- b) Map Reduce
- c) Hive
- d) All of the mentioned

Answer: c

26. _____ was designed to overcome limitations of the other Hive file formats.

- a) ORC
- b) OPC
- c) ODC
- d) None of the mentioned

Answer: a

27. An ORC file contains groups of row data called :

- a) postscript
- b) stripes
- c) script
- d) none of the mentioned

Answer: b

28. What are the types of hypervisor?

- There are two types of hypervisor
- Type1-hypervisor
- Type2-hypervisor

29. Define hybrid cloud?

It is a combination of private cloud and combined with the use of public cloud services with one or several touch points between environment.

30. What are the characteristics of bigdata lifecycle?

- capture
- organize
- integrate
- analyze
- act

UNIT-IV

31. _____ command is used to copy file or directories recursively.

- a) dtcp
- b) distcp
- c) dcp
- d) distc

Answer: b

32. R has how many atomic classes of objects ?

- a) 1
- b) 2
- c) 3
- d) 5

Answer: d

33. Numbers in R are generally treated as _____ precision real numbers.

- a) single
- b) double
- c) real
- d) all of the mentioned

Answer:b

34. If you explicitly want an integer, you need to specify the _____ suffix.

- a) D
- b) R
- c) L
- d) All of the mentioned

Answer:c

35. Attributes of an object (if any) can be accessed using the _____ function.

- a) objects()
- b) attrib()
- c) attributes()
- d) all of the mentioned

Answer:c

36. R objects can have attributes, which are like _____ for the object.

- a) metadata
- b) features
- c) expression
- d) none of the mentioned

Answer:a

37. Which of the following can be considered as object attribute ?

- a) dimensions
- b) class
- c) length
- d) all of the mentioned

Answer:d

38. The _____ function returns a list of all the formal arguments of a function

- a) formals()
- b) funct()

- c) formal()
- d) all of the mentioned

Answer:a

39. You can check to see whether an R object is NULL with the _____ function.

- a) is.null()
- b) is.nullobj()
- c) null()
- d) all of the mentioned

Answer:a

40. Which of the following code will print NULL ?

- a) > args(paste)
- b) > arg(paste)
- c) > args(pastebin)
- d) all of the mentioned

Answer:a

UNIT-V

41.What will be the output of following code snippet ?

```
> paste("a", "b", se = ":")
```

- a) "a+b"
- b) "a=b"
- c) "a b :"
- d) none of the mentioned

Answer:a

42. The _____ function returns a list of all the formal arguments of a function

- a) formals()
- b) funct()
- c) formal()
- d) all of the mentioned

Answer:a

43. Point out the wrong statement :

- a) A formal argument can be a symbol, a statement of the form 'symbol = expression', or the special formal argument
- b) The first component of the function declaration is the keyword function

- c) The value returned by the call to function is not a function
- d) None of the mentioned

Answer:a

44. You can check to see whether an R object is NULL with the _____ function.

- a) is.null()
- b) is.nullobj()
- c) null()
- d) all of the mentioned

Answer:a

45. Which of the following code will print NULL ?

- a) > args(paste)
- b) > arg(paste)
- c) > args(pastebin)
- d) all of the mentioned

Answer:a

46. Which of the following is apply function in R ?

- a) apply()
- b) tapply()
- c) fapply()
- d) all of the mentioned

Answer:b

47. Functions are defined using the _____ directive and are stored as R objects

- a) function()
- b) funct()
- c) functions()
- d) all of the mentioned

Answer:a

48. What will be the output of the following code ?

```
> f <- function() {  
+   ## This is an empty function  
+ }  
> f()
```

- a) 0
- b) No result

- c) NULL
- d) All of the mentioned

Answer:c

49. What will be the output of the following code ?

```
> x <- 3  
> switch(2, 2+2, mean(1:10), rnorm(5))
```

- a) 5
- b) 5.5
- c) NULL
- d) All of the mentioned

Answer:b

50. Which will be the output of following code ?

```
> x <- 3  
> switch(6, 2+2, mean(1:10), rnorm(5))
```

- a) 10
- b) 1
- c) NULL
- d) 5

Answer:c

NGM COLLEGE
PG DEPARTMENT OF COMPUTER SCIENCE
COURSE TITLE: BIG DATA ANALYTICS
COURSE CODE:17PCS313

K2 LEVEL QUESTIONS

UNIT-I

1. Mention what is the responsibility of a Data analyst?

- Provide support to all data analysis and coordinate with customers and staffs
- Resolve business associated issues for clients and performing audit on data

2. Mention what are the various steps in an analytics project?

Various steps in an analytics project include

- Problem definition
- Data exploration

3. List out some of the best practices for data cleaning?

Some of the best practices for data cleaning includes,

- Sort data by different attributes
- For large datasets cleanse it stepwise and improve the data with each step until you achieve a good data quality

4. List of some best tools that can be useful for data-analysis?

- RapidMiner
- OpenRefine
- Google Search Operators

5. List out some common problems faced by data analyst?

Some of the common problems faced by data analyst are

- Common misspelling
- Duplicate entries
- Missing values

6. Mention what are the data validation methods used by data analyst?

Methods used by data analyst for data validation are

- Data screening
- Data verification

7. Mention what are the key skills required for Data Analyst?

- Database knowledge
- Predictive Analytics
- Big Data Knowledge
- Presentation skill

8. Mention the tools used in Big Data?

Tools used in Big Data includes

- Hadoop
- Hive
- Pig
- Flume
- Mahout
- Sqoop

9. Explain what is Map Reduce?

Map-reduce is a framework to process large data sets, splitting them into subsets, processing each subset on a different server and then blending results obtained on each.

10. Explain what is Clustering?

Clustering is a classification method that is applied to data. Clustering algorithm divides a data set into natural groups or clusters.

UNIT-II

11. How is Hadoop different from other parallel computing systems?

Hadoop is a distributed file system, which lets you store and handle massive amount of data on a cloud of machines, handling data redundancy. Go through this HDFS content to know how the distributed file system works. The primary benefit is that since data is stored in several nodes, it is better to process it in distributed manner. Each node can process the data stored on it instead of spending time in moving it over the network.

12. Define DataNode and how does NameNode tackle DataNode failures?

DataNode stores data in HDFS; it is a node where actual data resides in the file system. Each datanode sends a heartbeat message to notify that it is alive. If the namenode does not receive a message from datanode for 10 minutes, it considers it to be dead or out of place, and starts replication of blocks that were hosted on that data node such that they are hosted on some other data node. A BlockReport contains list of all blocks on a DataNode. Now, the system starts to replicate what were stored in dead DataNode.

13. Write the difference between Map Side join and Reduce Side Join?

Map side Join at map side is performed data reaches the map. You need a strict structure for defining map side join. On the other hand, Reduce side Join (Repartitioned Join) is simpler than map side join since the input datasets need not be structured. However, it is less efficient as it will have to go through sort and shuffle phases, coming with network overheads.

14. How can you transfer data from Hive to HDFS?

By writing the query:

```
hive> insert overwrite directory '/' select * from emp;
```

15. List the five V's of Big Data?

- Volume
- Velocity
- Variety
- Veracity

- Value

16. Tell us how big data and Hadoop are related to each other.

Big data and Hadoop are almost synonyms terms. With the rise of big data, Hadoop, a framework that specializes in big data operations also became popular. The framework can be used by professionals to analyze big data and help businesses to make decisions.

17. How is big data analysis helpful in increasing business revenue?

Big data analysis has become very important for the businesses. It helps businesses to differentiate themselves from others and increase the revenue. Through predictive analytics, big data analytics provides businesses customized recommendations and suggestions.

18. Explain the steps to be followed to deploy a Big Data solution.

- Data Ingestion
- Data Storage
- Data Processing

19. How would you transform unstructured data into structured data?

Unstructured data is very common in big data. The unstructured data should be transformed into structured data to ensure proper data analysis. You can start answering the question by briefly differentiating between the two. Once done, you can now discuss the methods you use to transform one form to another. You might also share the real-world situation where you did it. If you have recently been graduated, then you can share information related to your academic projects.

20. How to recover a NameNode when it is down?

The following steps need to execute to make the Hadoop cluster up and running:

1. Use the FsImage which is file system metadata replica to start a new NameNode.
2. Configure the DataNodes and also the clients to make them acknowledge the newly started NameNode.
3. Once the new NameNode completes loading the last checkpoint FsImage which has received enough block reports from the DataNodes, it will start to serve the client.

UNIT-III

21. Explain some important features of Hadoop.

Hadoop supports the storage and processing of big data. It is the best solution for handling big data challenges. Some important features of Hadoop are –

- Open Source – Hadoop is an open source framework which means it is available free of cost. Also, the users are allowed to change the source code as per their requirements.
- Distributed Processing – Hadoop supports distributed processing of data i.e. faster processing. The data in Hadoop HDFS is stored in a distributed manner and MapReduce is responsible for the parallel processing of data.

22. List the different modes in which Hadoop run.

- Standalone (Local) Mode – By default, Hadoop runs in a local mode i.e. on a non-distributed, single node. This mode uses the local file system to perform input and output operation.
- Pseudo-Distributed Mode – In the pseudo-distributed mode, Hadoop runs on a single node just like the Standalone mode. In this mode, each daemon runs in a separate Java process

23. Name the core components of Hadoop.

- HDFS
- MapReduce Engine

24. Mention the feature of commodity hardware?

Commodity hardware is a low-cost system identified by less-availability and low-quality. The commodity hardware comprises of RAM as it performs a number of services that require RAM for the execution.

25. How is NFS different from HDFS?

There are a number of distributed file systems that work in their own way. NFS (Network File System) is one of the oldest and popular distributed file storage systems whereas HDFS (Hadoop Distributed File System) is the recently used and popular one to handle big data.

26. How do Hadoop MapReduce works?

There are two phases of MapReduce operation.

- Map phase – In this phase, the input data is split by map tasks. The map tasks run in parallel. These split data is used for analysis purpose.
- Reduce phase- In this phase, the similar split data is aggregated from the entire collection and shows the result.

27. How to restart all the daemons in Hadoop?

To restart all the daemons, it is required to stop all the daemons first. The Hadoop directory contains sbin directory that stores the script files to stop and start daemons in Hadoop.

28. Name the database used for Hadoop.

HBase

29. Name the language used for Hadoop.

Pig Latin

30. List the components used in Pig Latin execution environment.

- Local Node
- Hadoop node

UNIT-IV

31. What is R?

This should be an easy one for data science job applicants. R is an open-source language and environment for statistical computing and analysis, or for our purposes, data science.

32. Can you write and explain some of the most common syntax in R?

Some of the basic syntax for R that's used most often might include:

— as in many other languages, # can be used to introduce a line of comments. This tells the compiler not to process the line, so it can be used to make code more readable by reminding future inspectors what blocks of code are intended to do.

"" — quotes operate as one might expect; they denote a string data type in R.

<- — one of the quirks of R, the assignment operator is <- rather than the relatively more familiar use of =. This is an essential thing for those using R to know, so it would be good to display your knowledge of it if the question comes up.

\ — the backslash, or reverse virgule, is the escape character in R. An escape character is used to “escape” (or ignore) the special meaning of certain characters in R and, instead, treat them literally.

33. What are some advantages of R?

Its open-source nature. This qualifies as both an advantage and disadvantage for various reasons, but being open source means it's widely accessible, free to use, and extensible.

Its package ecosystem. The built-in functionality available via R packages means you don't have to spend a ton of time reinventing the wheel as a data scientist.

Its graphical and statistical aptitude. By many people's accounts, R's graphing capabilities are unmatched.

34. What are the disadvantages of R?

Memory and performance. In comparison to Python, R is often said to be the lesser language in terms of memory and performance. This is disputable, and many think it's no longer relevant as 64-bit systems dominate the marketplace.

Open source. Being open source has its disadvantages as well as its advantages. For one, there's no governing body managing R, so there's no single source for support or quality control. This also means that sometimes the packages developed for R are not the highest quality.

Security. R was not built with security in mind, so it must rely on external resources to mind these gaps.

35. How do you assign a variable in R?

```
myVar <- 15  
print(myVar)
```

36. Why use R?

This is a variant of the "advantages of R" question. Reasons to use R include its open-source nature and the fact that it's a versatile tool for statistical plotting, analysis, and portrayal. Don't be afraid to give some personal reasons as well. Maybe you simply love the assignment operator in R or feel that it's more elegant than other languages—but always remember to explicate.

37. Write a custom function in R

```
myFunction <- function(arg1, arg2, ... ){  
  statements  
  return(object)  
}
```

38. How do you import data in R?

```
yourRDateHere <- read.csv("Data.csv", header = TRUE)
```

39. How do you install a package in R?

There are many ways to install a package in R. Some even include using the GUI. We're coders, so we're not going to give those attention.

```
install.packages("package_name")
```

40. What is the use of with() in R?

We use the with() function to write simpler code by applying an expression to a data set. Its syntax looks like this:

```
with(randomDataSet, expression.test(sample))
```

UNIT-V

41. When is it appropriate to use mode()?

By default, **mode()** gets or sets the storage mode of an object. It's default usage is equivalent to storage.mode(). A sample usage:

```
x <- 1:25
mode(x)
[1] "numeric"
y <- "helloWorld"
mode(y)
[1] "character"
mode(state.name)
[1] "character"
```

42. When is it appropriate to use the which() function?

The which() function loops through a logical object until the condition returns TRUE and returns the index (position) of the element.

43. How do you concatenate strings in R?

Concatenating strings in R is less than intuitive. You don't use a . operator, nor a + operator, and forget about the & operator. In fact, you don't use an operator at all. Concatenating strings in R requires the use of the paste() function. Here's an example:

```
hello <- "Hello, "
world <- "World."
paste(hello, world)
```

44. How do you read a CSV file in R?

Use the read.csv() function.

```
yourRDateHere <- read.csv("Data.csv", header = TRUE)
```

45. Can you create an R decision tree?

A decision tree is a familiar graph for data scientists. It represents choices and results through the graphical form of a tree. To keep things simple, let's just go over the basics.

Install the party package to get started with making the tree.

```
install.packages("party")
```

46. Why is R useful for data science?

R turns otherwise hours of graphically intensive jobs into minutes and keystrokes. In reality, you probably wouldn't encounter the language of R outside the realm of data science or an adjacent field. It's great for linear modeling, nonlinear modeling, time-series analysis, plotting, clustering, and so much more.

Simply put, R is designed for data manipulation and visualization, so it's natural that it would be used for data science.

47. Describe how R can be used for predictive analysis

As a data manipulation and visualization tool, R can most definitely be used for predictive analytics. Using the same sort of decision tree we developed earlier, one could predict how many shootings might occur in 2019 in Boston. R as a whole provides numerous tools and packages for predictive modeling, so it's the right tool for a data scientist.

48. Name some functions available in "dplyr" package.

Functions in dplyr package:

- filter
- select
- mutate
- arrange
- count

49. How would you create a new R6 Class?

To first create an object template, which consists of the "Data Members" and "Class Functions" present in the class.

An R6 object template comprises of these parts ->

- Class Name
- Private Data Members
- Public Member Functions

50. How would you make multiple plots onto a single page in R?

If you want to plot 4 graphs onto the same pane, you can use the below command:

```
par(mfrow=c(2,2))
```

NGM COLLEGE
PG DEPARTMENT OF COMPUTER SCIENCE
COURSE TITLE: BIG DATA ANALYTICS
COURSE CODE:18PCS208

K3 Level Questions:

UNIT-I

1. How to manage the data? Explain.
2. Compare and Contrast structured and unstructured data with examples.
3. How distributed computing emerged? Discuss.
4. Discuss the basics of distributed computing.
5. Elaborate the concept of real-time and non-real time requirements of big data.
6. Explain Machine generated structured data with example.
7. Explain Machine generated unstructured data with example.
8. Explain Human generated structured data with example.
9. Explain Human generated unstructured data with example.
10. What is use of redundant physical infrastructure? Explain.

UNIT-II

11. Discuss the redundancies in physical infrastructure needed for bid data.
12. Explain the cycle of big data management.
13. Discuss performance matters of big data management.
14. Justify the usage of role of a CMS in big data management.
15. What are the features needed to consider when real time data are being used? Discuss.
16. Explain the components that need to be included for integration of data types into big data environment.
17. How to get performance right while using distributed computing? Explain.
18. Discuss the properties of operational databases.

19. Exemplify the three classes of tools in big data analytics.

20. Discuss the importance of virtualization to big data.

UNIT-III

21. Explain i) Server Virtualization ii) Application Virtualization.

22. Explain i) Network Virtualization ii) Process and Memory Virtualization.

23. How to manage virtualization with the Hypervisor.

24. Discuss the implementation of virtualization to work with big data.

25. Explain Cloud deployment models.

26. Explain Cloud delivery models.

27. Discuss the cloud characteristics to make an important part of big data ecosystem.

28. Discuss about the Amazon's public elastic compute cloud.

29. Exemplify the cloud services provided by Google.

30. Where to be careful when using cloud services? Explain.

UNIT-IV

31. Explain how you can create a table in R without external file?

32. Explain how to communicate the outputs of data analysis using R language.

33. Difference between library () and require () functions in R language.

34. How you can produce co-relations and covariances? Explain.

35. How many data structures R has? Explain.

36. Explain how data is aggregated in R.

37. Explain the purpose of using UIWindow object?

38. Compare R and Python programming languages for Predictive Modelling.

39. R language has several packages for solving a particular problem. How do you make a decision on which one is the best to use?

40. How will you create scatterplot matrices in R language? Explain.

UNIT-V

41. Explain how you can start the R commander GUI?
42. In R how can you can import Data?
43. Mention what does not 'R' language do?
44. Explain how R commands are written?
45. How can you save your data in R?
46. Mention how you can produce co-relations and covariances?
47. Explain what is With () and By () function in R is used for?
48. Explain general format of Matrices in R?
49. Explain what is transpose?
50. Explain how data is aggregated in R?

NGM COLLEGE
PG DEPARTMENT OF COMPUTER SCIENCE
COURSE TITLE: BIG DATA ANALYTICS
COURSE CODE:18PCS208

K4 & K5 Level Questions:

UNIT-I

1. Discuss about the waves of managing the big data.
2. Explain the architecture of Big Data with neat diagram.
3. What are sources of big data? Explain with its categories.
4. Elaborate Layer 0 of big data.
5. Elaborate Layer 1 of big data.

UNIT-II

6. Justify why interfaces and feeds to and from applications and the internet is necessary?
7. Explain Layer 2 of big data with its properties.
8. Elucidate the concept of implementing Virtualization in big data.
9. Explain Cloud deployment and delivery models.
10. Who are providers in the big data cloud market? Discuss.

UNIT-III

11. Explain Non-Relational databases used by big data.
12. Compare Non-relational databases with Document databases.
13. Explain i) Columnar database ii) Graph databases.
14. Is it necessary to put Map and Reduce function together? Discuss.
15. Why it is necessary to optimize MapReduce tasks? Explain.

UNIT-IV

16. Clarify how HDFS is essential for implementation of big data.

17. Explain the technology, language and database used for Hadoop.
18. Explain in detail about the benefits of using R with its unique features.
19. Explain the exploring RGui and dressing up with RStudio.
20. How to put argument in function? Explain.

UNIT-V

21. Discuss the naming conventions and code structure in R.
22. How clustering algorithm is implemented in R? Explain.
23. Perform all matrix calculations using R.
24. Write a program to plot a graph and normal distribution using R.
25. How Classification algorithm is implemented in R? Explain.