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

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 characterstics 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. _____ class allows the Map/Reduce framework to partition the map outputs based on certain key fields, not the whole keys.
a) KeyFieldPartitioner
b) KeyFieldBasedPartitioner
c) KeyFieldBased
d) None of the mentioned
Answer: b

33. Which of the following class is provided by Aggregate package ?
a) Map
b) Reducer
c) Reduce
d) None of the mentioned
Answer: b

34. An _____ is responsible for creating the input splits, and dividing them into records.
a) TextOutputFormat
b) TextInputFormat
c) OutputInputFormat
d) InputFormat
Answer: d

35. A _____ serves as the master and there is only one NameNode per cluster.
a) Data Node
b) NameNode
c) Data block
d) Replication
Answer: b

36. HDFS works in a _____ fashion.
a) master-worker
b) master-slave
c) worker/slave
d) all of the mentioned
Answer: a

37. What do you mean by commodity hardware?

Commodity Hardware refers to the minimal hardware resources needed to run the Apache Hadoop framework.

Any hardware that supports Hadoop's minimum requirements is known as 'Commodity Hardware.'

38.What is MapReduce? How does it work?

MapReduce is a parallel-programming model in the Hadoop file system. It is best for applications that must compute large data sets across computers.

MapReduce executes operations in two phases. In the Map phase, input is split into map tasks that work in tandem. In the Reduce phase, split data is brought back to provide overall output.

38.How Can You Transfer Data From Hive To HDFS?
Query:
Hive>insert overwrite directory '/' select * from emp;

39.What are the characteristics of operational database?

- Atomicity
- Consistency
- Isolation
- Durability

40.What are the pig mode?

- There are two mode
- local mode
- Hadoop

UNIT-V

41.Hive support language called
a)HiveQL
b)SQL
c)pigLatin
d)Noneoftheabove
Answer:a

42. _____ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.
a)Scalding
b)Casclog
c)Hcatalog
d)Hcalding
Answer:b

43. What is a unit of data that flows through a Flume agent?
a)Record
b)Event
c)Row
d)Log
Answer:b

44. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including:
a) Improved data storage and information retrieval
b) Improved extract, transform and load features for data integration
c) Improved data warehousing functionality
d) Improved security, workload management and SQL support
Answer:d

45. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop ?
a) Big data management and data mining
b) Data warehousing and business intelligence
c) Management of Hadoop clusters
d) Collecting and storing unstructured data
Answer:a

46. Hadoop clusters running on Amazon EMR use _____ instances as virtual Linux servers for the master and slave nodes
a) EC2
b) EC3
c) EC4
d) None of the mentioned
Answer:a

47. Impala executes SQL queries using a _____ engine.
a) MAP
b) MPP
c) MPA
d) None of the mentioned
Answer:b

48.What should be installed before using ZooKeeper?
Ans: Java

49. How many types of special znodes are present in Zookeeper ?
Ans: 2

50. Which was created to allow you to flow data from a source into your Hadoop environment?
Ans: Flume

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 noit 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. How does HDFS Index Data blocks?

HDFS indexes data blocks based on their respective sizes. The end of a data block points to the address of where the next chunk of data blocks gets stored.

32. Name a few data management tools used with Edge Nodes?

Oozie, Flume, Ambari, and Hue are some of the data management tools that work with edge nodes in Hadoop.

33. How many modes can Hadoop be run in?

Hadoop can be run in three modes— Standalone mode, Pseudo-distributed mode and fully-distributed mode.

34. Name the core methods of a reducer

The three core methods of a reducer are,

1. setup()
2. reduce()

3. cleanup()

35. Name a few companies that use Hadoop.

Yahoo, Facebook, Netflix, Amazon, and Twitter.

36. Name some Big Data Products?

- R
- Rattle
- Hadoop

37. Where does Big Data come from?

There are three sources of Big Data

- Social Data: It comes from social media channel's insights on consumer behaviour.
- Machine Data: It consists of real time data generated from sensors and web logs. It tracks user behaviour online.
- Transaction Data: It generated by large retailers and B2B Companies

38. Where the Mappers Intermediate data will be stored?

- The mapper output is stored in the local file system of each individual mapper node.
- Temporary directory location can be setup in configuration

39. How are file systems checked in HDFS?

- File system is used to control how data are stored and retrieved.
- Each file system have a different structure and logic properties of speed, security, flexibility, size.
- Such kind of file system designed in hardware. This file includes NTFS, UFS, XFS, HDFS.

40. Pig Latin contains different relational operations; name them?

- group
- distinct
- join
- for each

UNIT-V

41. Why are counters useful in Hadoop?

- Counter is an integral part of any Hadoop job.
- It is very useful gathering relevant statistics.
- Particular job consists of 150 node clusters with 150 mappers.

42. How are big data solutions developed?

Big data solutions are devised using a three-step process. First, data is ingested. Which means, it is gathered from a variety of sources like a CRM or log files like a social media feed. Data can be gathered, either in batches or in real-time by live streaming.

43. Can you mention a few problems that data analyst usually encounter while performing the analysis?

The following are a few problems that are usually encountered while performing data analysis.

- Presence of Duplicate entries and spelling mistakes, reduce data quality.
- If you are extracting data from a poor source, then this could be a problem as you would have to spend a lot of time cleaning the data.

44. Expand YARN.

Yet Another Resource Negotiator

45. Give the advantage of Zookeeper.

It can handle distributed applications effectively and efficiently.

46. Discuss the features of Zookeeper.

- Process Synchronization
- Configuration Management
- Self-Election
- Reliable Messages

47. List the keys skills a data analyst needs.

Predictive Analytics, Database Knowledge, Presentation Skills and Predictive Analytics.

48. Describe MapReduce.

This is a framework used for processing massive data sets, cutting them down into subsets then processing the subsets on a distinct server then the results obtained are blended.

49. Mention platforms which are used for large scale cloud computing?

The platforms that are used for large scale cloud computing are

a) Apache Hadoop

b) MapReduce

50. Mention some open source cloud computing platform databases?

The open source cloud computing platform databases are

a) MongoDB

b) CouchDB

c) LucidDB

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

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 PostgreSQL relational databases with example.

32. Demonstrate the concept of Key-Value pair databases with example.

33. Explain Riak key-value database with its features.

34. Compare and contrast Mongo and Couch Database.

35. Explain HBase columnar database with its characteristics.

36. Name the database used for graphical purpose. Explain.

37. Justify the usage of PostGIS/OpenGEO Suite in big data.

38. What is necessity of Polygot Persistence? Explain.

39. Explain the use of Map function.

40. Explain the use of reduce function.

UNIT-V

41. How to put map and reduce together? Explain.

42. Discuss the foundational behaviors of MapReduce.

43. Discuss the characteristics of file system in MapReduce.

44. Discuss the concept of HDFS with neat sketch.

45. Explain the use of namenode and datanode in HDFS.

46. With neat sketch explain workflow and data movement in a small Hadoop cluster.

47. How to mine big data with Hive? Explain.

48. Explain i) Sqoop ii) Zookeeper.

49. What are ten do's and don's of bigdata?

50. Elucidate the applications of Bigdata.

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

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 it 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. Exemplify various document databases.

19. Differentiate Mongo DB and Couch DB.

20. Elucidate the concept of Key-Value pair database.

UNIT-V

21. Explain any two applications of BigData Analytics.

22. How BigData is applied in Educational Institutions.

23. How to keep Data Analytics perspective? Exlpain.

24. Discuss the concept of Big Data Strategy with its planning.

25. Explain the transformation of Bigdata to Business.