P.G. Department of Computer Applications
NGM college, Pollachi

Big Data Management
III -MCA

K1 – type questions

Unit:1

1) **_____is a term associated with complex and large datasets.**
   a) Data warehouse b) Data mart c)Big data d) data management

2) **Data warehouse and_____ solved many problems for company needing a consistent way to manage massive transactional data.**
   a) Web b) Data mart c)Big data d) data management

3) **_____ was designed by google as a way of efficiently executing a set of functions against a large amount of data in batch mode.**
   a) Hadoop b) big table c) MapReduce d)data service

4) **_____ is an Apache-managed software framework derived from MapReduce and Big Table.**
   a) Hadoop b) big data c) data service d) data mart

5) **_____ is important because we are dealing with so much data from so many different sources.**
   a) Virtualization b) security infrastructure c) redundancy d) data source

6) **The data that created by a machine without human intervention is _____**
   a) Human-generated b) computer or machine generated
   c) sensor data d) financial data

7) **_____ refers to how a database retains versions of itself when modified.**
   a) Data persistence b) input data c) RDBMS d) point-of-sale data

8) **The great granddaddy of persistent data stores is the _____**
   a) RDBMS b) data persistent c)web log data d)financial data

9) **----------is a technique that allows individual computers to be networked together across geographical areas as though they were a single environment**
   a) Distributed computing b)RDBMS c)consistent models d)big data

10) **RPC _____**

   a) rapid producer call b)remote procedures calls

   c) rapid procedures calls d)remote producer call

Answers:

1) C) big data
2) B)data mart
3) c) MapReduce
4) a) Hadoop
5) c) redundancy
6) b) computer or machine generated
7) a) Data persistence
8) a) RDBMS
9) a) Distributed computing
10) b)remote procedures calls

Unit:2

1) Performance, also called _____ is often measured end to end, based on a single transaction or query request.
a)latency b)availability c)scalability d)cost

2) _____is the most challenging aspect of security in a big data environment.
a)Application access b)data access

c)threat detection d) Data encryption

3) _____(from transactions, applications, and so on) can produce a lot of disparate data.
a) Dozens b) Content sources c) Traditional data streams d) data warehouses

4) _____separates resources and services from the underlying physical delivery environment, enabling you to create many virtual systems within a single physical system.
a)encapsulation b) Virtualization c)partitioning d)isolation

5) _____can provide a level of automation and standardization to optimize your computing environment.
a)encapsulation b) Virtualization c)partitioning d)isolation

6) _____provides a foundation for cloud computing.
a)encapsulation b) Virtualization c)partitioning d)isolation

7) _____uses the hypervisor to provide efficiency in the use of physical resources.
a) Server virtualization b) Application infrastructure c) Network virtualization d) virtualization

8) _____virtualization used in combination with server virtualization can help to ensure that business service-level agreements (SLAs) are met.
a) Server virtualization b) Application infrastructure c) Network virtualization d) virtualization

9)_____software-defined networking provides an efficient way to use networking as a pool of connection resources.
a) Server virtualization b) Application infrastructure c) Network virtualization d) virtualization

10) The _____is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies and individuals.
a) public cloud b)private cloud
c)cloud computing d)distributed computing

Answers:
1) a)latency
2) d) Data encryption
3) c) Traditional data streams
4) b) Virtualization
5) b) Virtualization
6) b) Virtualization
7) a) Server virtualization
8) b) Application infrastructure
9) c) Network virtualization
10) a) public cloud

UNIT – 3

1. Key Value Pair Database do not require a schema (True/False)

   True
2. Name the data-interchange format, based on a subset of the JavaScript programming language.

   JSON
3. Give two examples of document database

   MongoDB

   CouchDB
4. Why we need polygotpersistence ?

   A polyglot persistence database is used when it is necessary to solve a complex problem by breaking that problem into segmentsand applying different database models.
5. Give an example for procedural programming model

   C++/FORTRAN/COBOL/C
6. What is the major difference between local and distributed file system

   Capacity
7. Classify the two primary components of Hadoop

   HDFS

   MapReduce Engine
8. Why HDFS supports data pipelines

   A pipeline is a connection between multiple data nodes that exists to support the movement of data across the servers in HDFS
9. Classify the pig execution environment modes

   Local Mode and Hadoop
10. What Sqoop represents?

    SQL-to-Hadoop

UNIT-4

1. Is it true that the data warehouse is subject oriented?

   Yes
2. What type of data is controlled by Hadoop based infrastructure?

   The data that is highly distributed and subject to change in real time
3. Loading information in a datawarehouse is the same as loading in big data?

No, both are entirely different

4. The cloud is becoming a compelling platform to manage big data. What does this mean?
   Since the volume of the data is increasing exponentially, Cloud becomes the compelling platform to manage big data
5. What does monetized analytics mean?
   Analytics which re utilized to directly drive revenue.
6. Give an example of advanced analytics
   Prediction, Forecasting and Complex Event Processing
7. State the concept in one word
   "Changing the internal code without affecting its external functioning"
   Refactoring
8. Give an example of big data analytics which uses predictive models to analyze safety data on aircrafts.
   NASA
9. What kind of data is collected by Flume
   Log Data
10. Is it valid that ETL has been used with batch processing in data warehouse environments.
    Yes

UNIT-5

1. Is "checking" a stage in business planning tool?
   Yes
2. Point out the correct statement :
   a) Hadoop do need specialized hardware to process the data
   b) Hadoop 2.0 allows live stream processing of real time data
   c) In Hadoop programming framework output files are divided in to lines or records
   d) None of the mentioned
   Answer – b
3. Facebook Tackles Big Data With _____ based on Hadoop.
   a) 'Project Prism'
   b) 'Prism'
   c) 'Project Big'
   d) 'Project Data'

Answer – a

4. _____ has the world's largest Hadoop cluster.

   a) Apple

   b) Datamatics

   c) Facebook

   d) None of the mentioned

   Answer – c

5. _____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.

   a) Reduce

   b) Map

   c) Reducer

   d) All of the mentioned

   Answer – a

6. The number of maps is usually driven by the total size of :

   a) inputs

   b) outputs

   c) tasks

   d) None of the mentioned

   Answer – a

7. Name one nonprofit organization that has started to focus on big data standards.

   OASIS

8. Write a technique for distributed computing that effectively manage the size, variety, and required speed to manage your data.

   Hadoop

9. Can we use third party data as our data source

   Yes, with permission

10. Can we leverage cloud computing services?

    yes.

P.G DEPARTMENT OF COMPUTER APPLICATIONS

K2 QUESTIONS WITH ANSWERS

BIG DATA MANAGEMENT - (17PMC639)

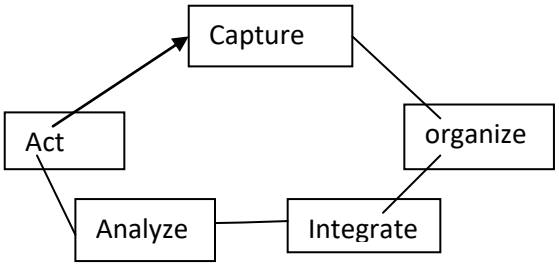
K1 TYPE QUESTIONS

UNIT- 1

1) Define big data.

   Big data is not a simple technology but a combination of old and new technologies that helps companies gain actionable insight. Therefore, big data is the capability to manage a huge volume of disparate data.

2) Discuss the cycle of bigdata management.



3) Discuss about security infrastructure.

   The more important big data analysis becomes to companies, the moreimportant it will be to secure that data. For example, if you are a healthcare company, you will probably want to use big data applications to determinechanges in demographics or shifts in patient needs.

4) Define hadoop.

   Hadoop is an Apache-managed software framework derived from MapReduceand Big Table. Hadoop allows applications based on MapReduce to run onlarge clusters of commodity hardware.

5) List out the categories of sources of data.

   1)Computer- or machine-generated: Machine-generated data generallyrefers to data that is created by a machine without human intervention.

   2)Human-generated: This is data that humans, in interaction with computers,Supply.

6) What are the machine generated instructured data?

   Sensor data

Web log data

Point-of-sale data

Financial data

7) What are the human generated instructured data?

Input data

Click-stream data

Gaming-related data

8) Define unstructured data.

Unstructured data is data that does not follow a specified format. If 20 percentof the data available to enterprises is structured data, the other 80 percentis unstructured. Unstructured data is really most of the data that you willencounter.

9) Why we need distributed computing for bigdata?

Not all problems require distributed computing. If a big time constraintdoesn't exist, complex processing can done via a specialized service remotely.When companies needed to do complex data analysis, IT would move datato an external service or entity where lots of spare resources were availablefor processing.

10) Define latency.

Latency is the delay within asystem based on delays in execution of a task. Latency is an issue in everyaspect of computing, including communications, data management, systemperformance, and more. If you have ever used a wireless phone, you haveexperienced latency firsthand.

UNIT -II

1) Expand ACID.

Atomicity Consistency Isolation Durability

2) List out the technologies used in organizing data services and tools.

A distributed file system, Serialization services, Coordination services,

Extract, transform, and load (ETL) tools, Workflow services

3) Define 3 classes of bigdata analytics.

Reporting and dashboards, Visualization, Analytics and advanced analytics

4) What is virtualization?

Virtualization separates resources and services from the underlying physicaldelivery environment, enabling you to create many virtual systems within asingle physical system.

5) Three categories in virtualization?

Partitioning, Isolation, Encapsulation

6) Define encapsulation.

A virtual machine can be represented (and even stored)as a single file, so you can identify it easily based on the services it provides.For example, the file containing the encapsulated process couldbe a complete business service.

7) Tell the benefits of virtualization.

- ✓ Virtualization of physical resources (such as servers, storage, andnetworks) enables substantial improvement in the utilization of theseresources.

✓Virtualization enables improved control over the usage and performance of your IT resources.

✓Virtualization can provide a level of automation and standardization to optimize your computing environment.

✓Virtualization provides a foundation for cloud computing.

8) Define server virtualization.

In server virtualization, one physical server is partitioned into multiple virtual servers. The hardware and resources of a machine — including the random access memory (RAM), CPU, hard drive, and network controller — can be virtualized (logically split) into a series of virtual machines that each runs its own applications and operating system.

9) What is cloud computing?

Cloud computing is a method of providing a set of shared computing resources that include applications, computing, storage, networking, development,and deployment platforms, as well as business processes. Cloudcomputing turns traditional siloed computing assets into shared pools ofresources based on an underlying Internet foundation.

10) Mention the deployment models for cloud computing.

A Public cloud, A Private cloud , A Hybrid cloud.

UNIT-III

1. Expand HDFS

   HadoopDistibuted File System
2. Expand CURD

   Create, Retrieve, Update and Delete
3. Name one widely used open source key-value pair database

   Riak
4. Write any two document database

   Mongo DB, Couch DB
5. Hbase belongs to which database?

   Columnar Database
6. Write an example of graph database

   Neo4J
7. Define Map Reduce

   MapReduce is a software framework that enables developers to write programs that can processmassive amounts of unstructured data in parallel across a distributed groupof processors.
8. What are the two primary components of Hadoop

   HDFS

   Map Reduce Engine
9. Expand YARN

   Yet Another Resource Negotiator
10. What are the services of YARN

    Global Resource Management

    Pre-Application Management

UNIT-IV

1. What is slicing and Dicing

   Slicing and dicing refers to breaking down your data into smaller sets of data that are easier to explore.

2. Define Predictive Modeling

   Predictive model is a statistical or data-mining solution consisting of algorithms and techniques that can beused on both structured and unstructured data (together or individually) to determine future outcomes.
3. Define Logistic Regression

   A statistical technique that is a variant of standard regression but extends the concept to deal with classification
4. What is Tableau?

   Tableau (www.tableausoftware.com), a business analytics and datavisualization software company, offers its visualization capabilities torun on top appliances and other infrastructure offered by a range of bigdata partners.
5. Define Pentaho

   Pentaho (www.pentaho.com) provides open source business analytics via a community and enterprise edition
6. Expand ETL

   Extraction, Transformation and Loading
7. What is FlumeNG

   FlumeNG loads data in real time by streaming the data into Hadoop.
8. Name the software vendors which provides ETL software tools

   IBM, Informatica,, Pervasive, Talend, and Pentaho,
9. Name the Hadoop tools for the transformation process

   HiveQL and Pig Latin
10. Name the strategy used by ELT to transform data

    SQL

UNIT-V

1. What are the stages of Big Data as Business Planning Tool?

   Planning, Doing, Checking and Acting
2. Write any two best practices of Big Data

   Understanding the goals

   Discover the data
3. Name any two organizations to work on big data

Open data foundation

Cloud Security Alliance

4. What is NIST

   NIST is a US government agency that focuses on emerging standard efforts

5. Expand OASIS

   Organization for the Advancement of Structured Information Standards

6. Name any four vendor sites for big data

   Google, IBM, Oracle, Microsoft

7. What are the New Dimensions to the Planning Cycle?

   Monitoring, Adjusting and Experimenting

8. Who sponsored Hadoop World

   O'Reilly

9. Write any two ideas to follow when we work with big data

   Involve all business units in the big data strategy

   Evaluate all delivery models for big data

10. Write any two ideas that could not be followed when we work with big data

    Don't overlook the need to integrate data

    Don't forget to manage data securely

P.G. Department of Computer Applications
NGM college, Pollachi

Big Data Management

K3 – type questions

Unit:1

1. Describe the cycle of Big data Management.
2. History of Distributed computing.
3. Various source of structured data.
4. Discuss web and content management.
5. Explain Real-time and non-Real-time requirements.
6. Discuss the integrating data types into a big data environment.
7. Explain about putting big data together.
8. Examine the role of relational databases in bigdata.
9. Analyze about traditional and advanced analytics.
10. Discuss about creating manageable data structures.
11. Unit:2
12. Classify the Big Data Analytics.
13. Classify the Big Data Applications.
14. Describe the basics of Virtualization.
15. Explain about processor and memory virtualization.
16. Explain about Data and Storage virtualization.
17. Analyze the virtualization with the hypervisor.
18. Discuss about cloud deployment model.
19. List out the cloud delivery models.
20. Contrast cloud as an Imperative for big data.
21. Classify the making use of the cloud for big data.

Unit – III

22. Narrate the charecteristics of non relational database
23. When do we use riak database
24. Discuss Hbase columnar database
25. Discuss polyglot persistence
26. Differentiate functional versus procedural programming model
27. Explain dataflow in Mapreduce with an example
28. How do you build a big data foundation with the hadoop ecossystem

29. How do you manage resources and applications with hadoop YARN
30. Explain the three mechanism used by Hive for data organization
31. How do you manage big data with Hive

Unit – IV

32. Discuss the changing role of datawarehouse
33. Discuss the deployment models in big data era
34. Explain data mining
35. Discuss any three big data analytics examples
36. How to identify the data that we need
37. How do you prioritize the big data quality
38. Discuss best practices for data integration in the big data world
39. Discuss traditional ETL
40. Explain FlumeNG for big data integration
41. Discuss the integration and incorporation stage

Unit – V

42. Discuss the new dimensions to the planning cycle
43. How do you keep data analytics in perspective
44. Discuss any 5 big data activities that we should do in big data implementation
45. Discuss any 5 big data activities that we should not do in big data implementation
46. Explain the standard organizations of big data
47. Discuss any 5 big data best practices

P.G. Department of Computer Applications
NGM college, Pollachi

Big Data Management

K4 and k5 – type questions

Unit :1

1) Explain the architecture of big data management.
2) Explain the source of Un structured data.
3) Explain the source of structured data.
4) Contrast the brief history of Distributed Computing.
5) Classify the waves of Managing Data.

Unit:2

1) Explore the Big data stack.
2) Explain the importance of virtualization of big data.
3) List the providers in the big data market.
4) Explain the cloud deployment and cloud delivery models.
5) Describe the redundant physical infrastructure.

Unit – III

1. Explain Document database in detail
2. Explain the optimizing of Map Reduce tasks
3. Explain Hadoop Distributed File System in detail
4. How do you interact with hadoop ecosystem. Explain
5. Explain the Map and Reduce function in detail

Unit – IV

1. Explain the integration of big data with the traditional data warehouse
2. How do you use big data to get results
3. Modify the business intelligence products to handle big data
4. Discuss big data analytics solutions
5. Elaborate the fundamentals of big data integeration

Unit – V

1. Explain big data as a business planning tool
2. Discuss the ten best practices of big data
3. Discuss ten great big data resources
4. Explain the ten big data activities that we have to do during the implementation of big data
5. Explain the ten big data activities that we should not do during the implementation of big data