DEPARTMENT OF INFORMATION TECHNOLOGY

BIG DATA ANALYTICS QUESTION BANK

EACH QUESTION CARRIES 1.5 MARKS

Id	**1**
Question	Facebook Tackles Big Data With _____ based on Hadoop
A	Project Prism
B	Prism
C	Project Data
D	Project Bid

Id	**2**
Question	What are the 3v's of Big Data?
A	Volume
B	Variety
C	Velocity
D	all the above

Id	**3**
Question	What license is Hadoop distributed under ?
A	Apache License 2.0
B	Mozilla
C	Shareware
D	Middleware

Id	**4**
Question	Which of the following genres does Hadoop produce ?
A	Distributed file system
B	JAX-RS
C	Java Message Service
D	JSP

Id	**5**
Question	What was Hadoop written in ?
A	C
B	C++
C	Java
D	JSP

Id	**6**
Question	Which of the following platforms does Hadoop run on ?
A	Bare metal
B	Debian
C	Cross-platform
D	Unix-Like

Id	**7**
Question	Hadoop achieves reliability by replicating the data across multiple hosts, and hence does not require_____ storage on hosts.
A	RAID
B	ZFS
C	Operating System
D	DFS

Id	**8**
Question	Above the file systems comes the _____ engine, which consists of one Job Tracker, to which client applications submit MapReduce jobs.
A	MapReduce
B	Google
C	Functional Programming
D	Facebook

Id	**9**
Question	The Hadoop list includes the HBase database, the Apache Mahout _____ system, and matrix operations.
A	Machine learning
B	Pattern recognition
C	Statistical classification
D	Artificial intelligence

Id	**10**
Question	_____ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.
A	Pig Latin
B	Oozie
C	Pig
D	Hive

Id	**11**
Question	Point out the correct statement
A	Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data
B	Hive is a relational database with SQL support
C	Pig is a relational database with SQL support
D	All of the mentioned

Id	**12**
Question	_____ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.
A	Scalding
B	HCatalog
C	Cascalog
D	All of the mentioned

Id	**13**
Question	Hive also support custom extensions written in :
A	C
B	C++
C	C#
D	Java

Id	**14**
Question	Point out the wrong statement
A	Amazon Web Service Elastic MapReduce (EMR) is Amazon packaged Hadoop offering
B	Elastic MapReduce (EMR) is Facebook packaged Hadoop offering
C	Scalding is a Scala API on top of Cascading that removes most Java boilerplate
D	All of the mentioned

Id	**15**
Question	_____ is general-purpose computing model and run time system for distributed data analytics.
A	Mapreduce
B	Drill
C	Oozie
D	None of the mentioned

Id	**16**
Question	The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to :
A	JSON
B	XML
C	XSL
D	SQL

Id	**17**
Question	_____ jobs are optimized for scalability but not latency
A	Mapreduce
B	Drill
C	Hive
D	Chuckro

Id	**18**
Question	_____ is a framework for performing remote procedure calls and data serialization.
A	Mapreduce
B	Dril
C	Avro
D	Chuckro

Id	**19**
Question	Which of the following is not a collision resolution technique?
A	Separate chaining
B	Linear probing
C	Quadratic probing
D	Hashing

Id	**20**
Question	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

Id	**21**
Question	According to analysts, for what can traditional IT systems provide a foundation when they are 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

Id	**22**
Question	Hadoop is a framework that works with a variety of related tools. Common cohorts include
A	MapReduce, MySQL and Google Apps
B	MapReduce, Hive and HBase
C	MapReduce, Hummer and Iguana
D	MapReduce, Heron and Trumpet

Id	**23**
Question	Which of the following is not an input format in Hadoop ?
A	Text Input Format
B	Byte Input Format
C	Sequence File Input format
D	Kep Input Format

Id	**24**
Question	What was Hadoop named after?
A	Creator Doug Cutting favorite circus act
B	Cutting high school rock band
C	The toy elephant of Cutting son
D	A sound Cutting laptop made during Hadoop development

Id	**25**
Question	All of the following accurately describe Hadoop, EXCEPT
A	Open source
B	Real-time
C	Java-based
D	Distributed computing approach

Id	**26**
Question	A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the Job Tracker.
A	MapReduce
B	Mapper
C	TaskTracker
D	JobTracker

Id	**27**
Question	Point out the correct statement
A	Map Task in MapReduce is performed using the Mapper() function
B	Reduce Task in MapReduce is performed using the Map() function
C	MapReduce tries to place the data and the compute as close as possible All
D	of the mentioned

Id	**28**
Question	_____ part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.
A	Maptask
B	Mapper
C	Task execution
D	All of the mentioned

Id	**29**
Question	_____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.
A	Map
B	Reduce
C	Reducer
D	Reduced

Id	**30**
Question	Point out the wrong statement A.
A	MapReduce job usually splits the input data-set into independent chunks which are processed bythe map tasks in a completely parallel manner
B	The MapReduce framework operates exclusively on pairs
C	Applications typically implement the Mapper and Reducer interfaces to provide the map and reducemethods
D	None of the mentioned

Id	**31**
Question	_____ is a utility which allows users to create and run jobs with any executables as the mapper and/or the reducer.
A	HadoopStrdata
B	Hadoop Streaming
C	Hadoop Stream
D	None of the mentioned

Id	**32**
Question	_____ maps input key/value pairs to a set of intermediate key/value pairs.
A	Mapper
B	Reducer
C	Both Mapper and Reducer
D	None of the mentioned

Id	**33**
Question	The number of maps is usually driven by the total size of
A	task
B	output
C	input
D	none

Id	**34**
Question	_____ is the default Partitioner for partitioning key space
A	HashPar
B	Partitioner
C	HashPartitioner
D	None of the mentioned

Id	**35**
Question	Point out the correct statement
A	Applications can use the Reporter to report progress
B	The Hadoop MapReduce framework spawns one map task for each Input Split generated by the Input Format for the job
C	The intermediate, sorted outputs are always stored in a simple (key-len, key, value-len, value) format
D	All of the mentioned

Id	**36**
Question	Input to the _____ is the sorted output of the mappers.
A	Reducer
B	Mapper
C	Shuffle
D	All of the mentioned

Id	**37**
Question	The right number of reduces seems to be :
A	0.65
B	0.55
C	0.95
D	0.68

Id	**38**
Question	Point out the wrong statement
A	Reducer has 2 primary phases
B	Increasing the number of reduces increases the framework overhead, but increases load balancing and lowers the cost of failures
C	It is legal to set the number of reduce-tasks to zero if no reduction is desired
D	The framework groups Reducer inputs by keys (since different mappers may have output the same key) in sort stage

Id	**39**
Question	The output of the _____ is not sorted in the Mapreduce framework for Hadoop.
A	Mapper
B	Cascader
C	Scalding
D	None of the mentioned

Id	**40**
Question	Mapper and Reducer implementations can use the _____ to report progress or just indicate that they are alive.
A	Partitioner
B	OutputCollector
C	Reporter
D	All of the mentioned

Id	**41**
Question	_____ is a generalization of the facility provided by the MapReduce framework to collect data output by the Mapper or the Reducer
A	Partitioner
B	OutputCollector
C	Reporter
D	All of the mentioned

Id	**42**
Question	A _____ serves as the master and there is only one NameNode per cluster
A	Data Node
B	NameNode
C	Data block
D	Deplication

Id	**43**
Question	Point out the correct statement
A	DataNode is the slave/worker node and holds the user data in the form of Data Blocks
B	Each incoming file is broken into 32 MB by default
C	Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance
D	None of the mentioned

Id	**44**
Question	HDFS works in a _____ fashion
A	master-worker
B	master-slave
C	worker/slave.
D	All of the mentioned

Id	**45**
Question	_____ NameNode is used when the Primary NameNode goes down.
A	Rack
B	Data
C	Secondary
D	None

Id	**46**
Question	Point out the wrong statement
A	Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level
B	Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode
C	User data is stored on the local file system of DataNodes
D	DataNode is aware of the files to which the blocks stored on it belong to

Id	**47**
Question	Which of the following scenario may not be a good fit for HDFS?
A	HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
B	HDFS is suitable for storing data related to applications requiring low latency data access
C	HDFS is suitable for storing data related to applications requiring high latency data access
D	None of the mentioned

Id	**48**
Question	The need for data replication can arise in various scenarios like :
A	Replication Factor is changed
B	DataNode goes down
C	Data Blocks get corrupted
D	All of the mentioned

Id	**49**
Question	_____ is the slave/worker node and holds the user data in the form of Data Blocks
A	DataNode
B	NameNode
C	Data block
D	Replication

Id	**50**
Question	HDFS provides a command line interface called _____ used to interact with HDFS.
A	HDFS Shell
B	FS Shell
C	DFSA Shell
D	None

Id	**51**
Question	HDFS is implemented in _____ programming language
A	C++
B	Java
C	Scala
D	None

Id	**52**
Question	Point out the correct statement
A	Cloudera is also a sponsor of the Apache Software Foundation
B	CDH is 100% Apache-licensed open source and is the only Hadoop solution to offer unified batchprocessing, interactive SQL, and interactive search, and role-based access controls
C	More enterprises have downloaded CDH than all other such distributions combined
D	All of the mentioned

Id	**53**
Question	_____ is a online NoSQL developed by Cloudera.
A	HCatalog
B	Hbase
C	Imphala
D	Oozie

Id	**54**
Question	Point out the wrong statement
A	To read data from an HBase table, use the get() method of the HTable class
B	You can retrieve data from the HBase table using the get() method of the HTable class
C	While retrieving data, you can get a single row by id, or get a set of rows by a set of row ids, or scanan entire table or a subset of rows
D	None of the mentioned

Id	**55**
Question	_____ class adds HBase configuration files to its object.
A	Configuration
B	Collector
C	Component
D	None of the mentioned

Id	**56**
Question	HBase uses the _____ File System to store its data.
A	Hive
B	Impala
C	Hadoop
D	Scala

Id	**57**
Question	Which of the following is a principle of analytic graphics?
A	Don't plot more than two variables at at time
B	Make judicious use of color in your scatterplots
C	Show box plots (univariate summaries)
D	Show causality, mechanism, explanation

Id	**58**
Question	MapReduce was devised by_____
A	Apple
B	Google
C	Facebook
D	Samsung

Id	**59**
Question	_____ is the world's largest Hadoop Cluster.
A	Apple
B	Datamatics
C	Facebook
D	None of the above

Id	**60**
Question	What are the main components of Big Data?
A	MapReduce
B	HDFS
C	YARN
D	all the above

Id	**61**
Question	What are the different features of Big Data Analytics?
A	Open Source
B	Data Recovery
C	Scalability
D	all of the above

Id	**62**
Question	For YARN, the _____ Manager UI provides host and port information.
A	Data Node
B	NameNode
C	Resource
D	Replication

Id	**63**
Question	Point out the correct statement
A	The Hadoop framework publishes the job flow status to an internally running web server on the master nodes of the Hadoop cluster
B	Each incoming file is broken into 32 MB by default
C	Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance
D	None of the mentioned

Id	**64**
Question	For _____, the HBase Master UI provides information about the HBase Master up time.
A	Oozie
B	HBase
C	Kafka
D	Afka

Id	**65**
Question	_____ Manager's Service feature monitors dozens of service health and performance metrics about the services and role instances running on your cluster.
A	Microsoft
B	Cloudera
C	Amazon
D	None of the abovc

Id	**66**
Question	NameNode is monitored and upgraded in a _____ transition.
A	safemode
B	securemode
C	servicemode
D	servicemonitor

Id	**67**
Question	HBase is a distributed _____ database built on top of the Hadoop file system.
A	Row-oriented
B	Column-oriented
C	Tuple-oriented
D	None of the mentioned

Id	**68**
Question	Kafka is run as a cluster comprised of one or more servers each of which is called_____
A	cTakes
B	Chunks
C	Broker
D	None of the mentioned

Id	**69**
Question	True or False ? Statement 1: Batch Processing provides ability to process and analyze data at-rest (stored data) Statement 2: Stream Processing provides ability to ingest, process and analyze data in-motion in real or near-real-time.
A	Only statement 1 is true
B	Only statement 2 is true
C	Both statements are true
D	Both statements are false

Id	**70**
Question	What are the parameters defined to specify window operation ? Window length, sliding interval
A	Window length, sliding interval
B	State size, window length
C	State size, sliding interval
D	None of the mentioned

Id	**71**
Question	Identify the correct choices for the given scenarios: P: The system allows operations all the time, and operations return quickly Q: All nodes see same data at any time, or reads return latest written value by any client R: The system continues to work in spite of network partitions
A	P: Consistency, Q: Availability, R: Partition tolerance
B	P: Availability, Q: Consistency, R: Partition tolerance
C	P: Partition tolerance, Q: Consistency, R: Availability
D	P: Consistency, Q: Partition tolerance, R: Availability

Id	**72**
Question	Consider the following statements: Statement 1: When two processes are competing with each other causing data corruption, it is called deadlock. Statement 2: When two processes are waiting for each other directly or indirectly, it is called race condition.
A	Only statement 1 is true
B	Only statement 2 is true
C	Both statements are true
D	Both statements are false

Id	**73**
Question	Consider the following statements in the context of Spark: Statement 1: Spark also gives you control over how you can partition your Resilient Distributed Datasets (RDDs) Statement 2: Spark allows you to choose whether you want to persist Resilient Distributed Dataset (RDD) onto disk or not.
A	Only statement 1 is true
B	Only statement 2 is true
C	Both statements are true
D	Both statements are false

Id	**74**
Question	Which of the following is not a NoSQL database?
A	HBase
B	SQL Server
C	Cassandra
D	None of the mentioned

Id	**75**
Question	Which of the following are the simplest NoSQL databases ?
A	Key-value
B	Wide-column
C	Document
D	All of the mentioned

Id	76
Question	Consider the following statements: Statement 1: The Job Tracker is hosted inside the master and it receives the job execution request from the client. Statement 2: Task tracker is the MapReduce component on the slave machine as there are multiple slave machines.
A	Only statement 1 is true
B	Only statement 2 is true
C	Both statements are true
D	Both statements are false

Id	77
Question	What should be the load factor for separate chaining hashing?
A	0.5
B	1
C	1.5
D	2

Id	78
Question	_____function processes a key/value pair to generate a set of intermediate key/value pairs.
A	Map
B	Reduce
C	Both Map and Reduce
D	None of the mentioned

Id	79
Question	True or False? The main duties of task tracker are to break down the receive job that is big computations in small parts, allocate the partial computations that is tasks to the slave nodes monitoring the progress and report of task execution from the slave.
A	TRUE
B	FALSE
C	SOMETIMES TRUE
D	SOMETIMES FALSE

Id	80
Question	Point out the correct statement in context of YARN:
A	YARN extends the power of Hadoop to incumbent and new technologies found within the data center.
B	YARN is highly scalable
C	YARN enhances a Hadoop compute cluster in many ways
D	All of the mentioned

Id	81
Question	Apache Hadoop YARN stands for:
A	Yet Another Reserve Negotiator
B	Yet Another Resource Network
C	Yet Another Resource Negotiator
D	Yet Another Resource Manager

Id	82
Question	Consider the pseudo-code for MapReduce's WordCount example. Let's now assume that you want to determine the frequency of phrases consisting of 3 words each instead of determining the frequency of single words. Which part of the pseudo-code do you need to adapt?
A	Only map()
B	Only reduce()
C	Both map() and reduce()
D	The code does not have to be changed

Id	83
Question	For which of the following operations is NO communication with the NameNode required?
A	A client writing a file to HDFS.
B	A client requesting the filename of a given block of data.
C	A client reading a block of data from the cluster.
D	A client reading a file from the cluster.

Id	84
Question	Which of the following components reside on a NameNode?
A	Filenames, blocks and checksums
B	Blocks and heartbeat messages
C	Filenames, block locations
D	Blocks and block locations

Id	85
Question	_____refers to the biases, noise and abnormality in data, trustworthiness of data.
A	Value
B	Veracity
C	Velocity
D	Volume

Id	86
Question	_____refers to the connectedness of big data.
A	Value
B	Veracity
C	Velocity
D	Valence

Id	87
Question	Which of the following is guaranteed by Zookeeper?
A	Interactivity
B	Flexibility
C	Scalability
D	Reliability

Id	88
Question	_____ is a distributed machine learning framework on top of Spark.
A	MLlib
B	Spark Streaming
C	GraphX
D	RDDs

Id	89
Question	_____is a resource management platform responsible for managing compute resources in the cluster and using them in order to schedule users and applications.
A	Hadoop Common
B	Hadoop Distributed File System (HDFS)
C	Hadoop YARN
D	Hadoop MapReduce

Id	90
Question	Which of the following tool is designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.
A	Apache Sqoop
B	Pig
C	Mahout
D	Flume

Id	91
Question	_____is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
A	Apache Sqoop
B	Pig
C	Mahout
D	Flume

Id	92
Question	_____ brings scalable parallel database technology to Hadoop and allows users to submit low latencies queries to the data that's stored within the HDFS or the Hbase without acquiring a ton of data movement and manipulation.
A	Apache Sqoop
B	Impala
C	Mahout
D	Flume

Id	93
Question	Which of the following is a characteristic of Big Data?
A	Huge volume of data
B	Complexity of data types and structures
C	Speed of data creation and growth
D	All of the mentioned

Id	94
Question	Concurrent access to shared data may result in _____
A	Data consistency
B	Data insecurity
C	Data inconsistency
D	None of the mentioned

Id	95
Question	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 into lines or records
D	None of the mentioned

Id	96
Question	Who will initiate the mapper?
A	Task tracker
B	Job tracker
C	Combiner
D	Reducer

Id	97
Question	Which of the following are the Big Data Solutions Candidates?
A	Processing 1.5 TB data everyday
B	Processing 30 minutes Flight sensor data
C	Interconnecting 50K data points (approx. 1 MB input file)
D	All of the above

Id	98
Question	Hadoop is a framework that allows the distributed processing of:
A	Small Data Sets
B	Semi-Large Data Sets
C	Large Data Sets
D	Large and Small Data sets

Id	99
Question	Which of the following is true about Name Node?
A	It is the Master Machine of the Cluster
B	It is Name Node that can store user data
C	Name Node is a storage heavy machine
D	Name Node can be replaced by any Data Node Machine

Id	100
Question	Which of the following are NOT metadata items?
A	List of HDFS files
B	HDFS block locations
C	Replication factor of files
D	File Records distribution

Id	101
Question	Name Node monitors block replication process.
A	TRUE
B	FALSE
C	Depends on file type
D	Vary depending on cluster

Id	102
Question	Which of the following are true for Hadoop Pseudo Distributed Mode?
A	It runs on multiple machines
B	Runs on multiple machines without any daemons
C	Runs on Single Machine with all daemons
D	Runs on Single Machine without all daemons

Id	103
Question	Which of the following is the highest level of Data Model in Hive?
A	Table
B	View
C	Database
D	Partitions

Id	104
Question	Snapshot in GFS creates a copy of a file or a directory tree at low cost
A	TRUE
B	FALSE
C	NOT ALWAYS
D	CAN'T SAY

Id	105
Question	The mechanism used to create replica in HDFS is_____.
A	Gossip protocol
B	Replicate protocol
C	HDFS protocol
D	Store and Forward protocol

Id	106
Question	In GFS Chunk size used is
A	21 MB
B	42 MB
C	64 MB
D	84 MB

Id	107
Question	NameNodes are usually high storage machines in the clusters.
A	TRUE
B	FALSE
C	Depends on cluster size
D	True if co-located with Job tracker

Id	108
Question	From the options listed below, select the suitable data sources for flume.
A	Publicly open web sites
B	Local data folders
C	Remote web servers
D	Both (A) and (C)

Id	109
Question	What should be an upper limit for counters of a Map Reduce job?
A	~5
B	~15
C	~150
D	~50

Id	110
Question	Which of the following is/are true about combiners?
A	Combiners can be used for mapper only job
B	Combiners can be used for any Map Reduce operation
C	Mappers can be used as a combiner class
D	Combiners are primarily aimed to improve Map Reduce performance

Id	111
Question	Pig is a:
A	Programming Language
B	Data Flow Language
C	Query Language
D	Database

Id	112
Question	What does commodity Hardware in Hadoop world mean?
A	Very cheap hardware
B	Industry standard hardware
C	Discarded hardware
D	Low specifications Industry grade hardware

Id	113
Question	What does “Velocity” in Big Data mean?
A	Speed of input data generation
B	Speed of individual machine processors
C	Speed of ONLY storing data
D	Speed of storing and processing data

Id	114
Question	Which of the following are example(s) of Real Time Big Data Processing?
A	Complex Event Processing (CEP) platforms
B	Stock market data analysis
C	Bank fraud transactions detection
D	Both (A) and (C)

Id	115
Question	What is HBase used as?
A	Tool for Random and Fast Read/Write operations in Hadoop
B	Faster Read only query engine in Hadoop
C	MapReduce alternative in Hadoop
D	Fast MapReduce layer in Hadoop

Id	116
Question	Which of the following are NOT true for Hadoop?
A	It’s a tool for Big Data analysis
B	It supports structured and unstructured data analysis
C	It aims for vertical scaling out/in scenarios
D	Both (A) and (C)

Id	117
Question	Which of the following are the core components of Hadoop?
A	HDFS
B	Map Reduce
C	HBase
D	Both (A) and (C)

Id	118
Question	Hive can be used for real time queries.
A	TRUE
B	FALSE
C	True if data set is small
D	True for some distributions

Id	119
Question	What is the default HDFS block size?
A	32 MB
B	32 KB
C	128 MB
D	64 MB

Id	120
Question	What is the default HDFS replication factor?
A	4
B	1
C	3
D	2

Id	121
Question	Which of the following is NOT a type of metadata in NameNode?
A	List of files
B	Block locations of files
C	Number of file records
D	File access control information

Id	122
Question	Which of the following is the correct sequence of MapReduce flow?
A	Combine ◊ Reduce ◊ Map
B	Map ◊ Combine ◊ Reduce
C	Reduce ◊ Combine ◊ Map
D	Map ◊ Reduce ◊ Combine

Id	123
Question	Which of the following can be used to control the number of part files in a map reduce program output directory?
A	Number of Mappers
B	Number of Reducers
C	Counter
D	Partitioner

Id	124
Question	Distributed Cache can be used in
A	Mapper phase only
B	Reducer phase only
C	In either phase, but not on both sides simultaneously
D	In either phase

Id	125
Question	What is optimal size of a file for distributed cache?
A	<=10 MB
B	>=250 MB
C	<=100 MB
D	<=35 MB

Id	126
Question	Number of mappers is decided by the
A	Mappers specified by the programmer
B	Available Mapper slots
C	Available heap memory
D	Input Splits

Id	127
Question	Distributed cache files can’t be accessed in Reducer
A	TRUE
B	FALSE
C	True if data set is small
D	True for some distributions

Id	128
Question	A Map reduce job can be written in:
A	Java
B	Ruby
C	Python
D	Any Language which can read from input stream

Id	129
Question	Just collecting and storing information isn't enough to produce real business value. Big data analytics technologies are necessary to:
A	Formulate eye-catching charts and graphs
B	Extract valuable insights from the data
C	Integrate data from internal and external sources
D	None of these

Id	130
Question	The method by which companies analyze customer data or other types of information in an effort to identify patterns and discover relationships between different data elements is often referred to as:
A	Data mining
B	Data digging
C	Customer data management
D	Data warehousing

Id	131
Question	True or false? To maximize the benefits of big data analytics techniques, it's critical for organizations to select the right tools and involve people who bring needed analytical skills to a project.
A	TRUE
B	FALSE
C	NOT ALWAYS
D	NEVER

Id	132
Question	What is the recommended best practice for managing big data analytics programs?
A	Adopting data analysis tools based on a laundry list of their capabilities
B	Letting go entirely of "old ideas" related to data management
C	Focusing on business goals and how to use big data analytics technologies to meet them
D	None of these

Id	133
Question	True or false? A big data analytics strategy is often defined by the three V's -- volume, variety and velocity -- which is helpful but ignores other commonly cited characteristics, such as complexity and variability.
A	TRUE
B	FALSE
C	CAN'T SAY
D	NOT ALWAYS

Id	134
Question	Companies that have large amounts of information stored in different systems should begin a big data analytics project by considering:
A	The creation of a plan for choosing and implementing big data infrastructure technologies
B	The interrelatedness of data and the amount of development work that will be needed to link various data sources
C	The ability of business intelligence and analytics vendors to help them answer business questions in big data environments
D	Data warehousing

Id	135
Question	True or false? For organizations that aren't currently looking to do big data analytics, there is little or no benefit to examining the data they're retaining and evaluating how it's being used.
A	TRUE
B	FALSE
C	CAN'T SAY
D	NOT ALWAYS

Id	136
Question	What is the name of the programming framework originally developed by Google that supports the development of applications for processing large data sets in a distributed computing environment?
A	MapReduce
B	Hive
C	ZooKeeper
D	DOS

Id	137
Question	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

Id	138
Question	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 into lines or records
D	None of the mentioned

Id	139
Question	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

Id	140
Question	_____ is general-purpose computing model and run time system for distributed data analytics.
A	Mapreduce
B	Drill
C	Oozie
D	None of the mentioned

Id	141
Question	Which of the following scenario may not be a good fit for HDFS?
A	HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
B	HDFS is suitable for storing data related to applications requiring low latency data access
C	HDFS is suitable for storing data related to applications requiring low latency data access
D	None of the mentioned

Id	142
Question	The need for data replication can arise in various scenarios like _____
A	Replication Factor is changed
B	DataNode goes down
C	Data Blocks get corrupted
D	All of the mentioned

Id	143
Question	In a hash table of size 10, where is element 7 placed?
A	6
B	7
C	17
D	16

Id	144
Question	Point out the wrong statement.
A	Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level
B	Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode
C	User data is stored on the local file system of DataNodes
D	DataNode is aware of the files to which the blocks stored on it belong to

Id	145
Question	What is a hash table?
A	A structure that maps values to keys
B	A structure that maps keys to values
C	A structure used for storage
D	A structure used to implement stack and queue

Id	146
Question	If several elements are competing for the same bucket in the hash table, what is it called?
A	Diffusion
B	Replication
C	Collision
D	Duplication

Id	147
Question	What is a hash function?
A	A function has allocated memory to keys
B	A function that computes the location of the key in the array
C	A function that creates an array
D	A function that computes the location of the values in the array

Id	148
Question	What is the load factor?
A	Average array size
B	Average key size
C	Average chain length
D	Average hash table length

Id	149
Question	What is simple uniform hashing?
A	Every element has equal probability of hashing into any of the slots
B	A weighted probabilistic method is used to hash elements into the slots
C	Elements has Random probability of hashing into array slots
D	Elements are hashed based on priority

Id	150
Question	In simple chaining, what data structure is appropriate?
A	Singly linked list
B	Doubly linked list
C	Circular linked list
D	Binary trees