

X



(<https://swayam.gov.in>)



(https://swayam.gov.in/nc_details/NPTEL)

remeshbabu@gecspk.ac.in ▾

NPTEL (<https://swayam.gov.in/explorer?ncCode=NPTEL>) » Big Data Computing (course)

≡

Register for
Certification
exam

(<https://examform.nptel.ac.in/>)

Course
outline

How does an
NPTEL online
course work?

Week-0

Quiz: Week-0:
Assignment-0
(assessment?
name=91)

Week-1

Week-0: Assignment-0

Your last recorded submission was on 2021-08-23, 15:56 Due date: 2021-08-23, 23:59 IST.

- 1) The maximum number of super keys for the relation schema R(E,F,G,H) with E as **1 point** the key is

- 5
- 6
- 7
- 8

- 2) Consider the following relational schemas for a library database: **1 point**

Book (Title, Author, Catalog_no, Publisher, Year, Price)

Collection (Title, Author, Catalog_no)

with in the following functional dependencies:

- I. Title, Author --> Catalog_no
- II. Catalog_no --> Title, Author, Publisher, Year
- III. Publisher, Title, Year --> Price

Assume {Author, Title} is the key for both schemas. Which of the following statements is true ?

- Both Book and Collection are in BCNF
- Both Book and Collection are in 3NF only
- Book is in 2NF and Collection is in 3NF
- Both Book and Collection are in 2NF only

- 3) Consider a B+-tree in which the maximum number of keys in a node is 5. What is **1 point** the minimum number of keys in any non-root node ?

- 1
- 2
- 3
- 4

4) Consider the join of a relation R, with a relation S. If R has m number of tuples and **1 point** S has n number of tuples then the maximum and minimum sizes of the join respectively are:

- $m + n \& 0$
- $mn \& 0$
- $m + n \& |m - n|$
- $mn \& m + n$

5) Which one of the following is NOT a part of the ACID properties of database **1 point** transactions ?

- Atomicity
- Consistency
- Isolation
- Deadlock-freedom

6) In the IPv4 addressing format, the number of networks allowed under Class C **1 point** addresses is:

- 2^{14}
- 2^7
- 2^{21}
- 2^{24}

7) One of the header fields in an IP datagram is the Time to Live (TTL) field. Which of **1 point** the following statements best explains the need for this field ?

- It can be used to prioritize packets
- It can be used to reduce delays
- It can be used to optimize throughput
- It can be used to prevent packet looping

8) The address resolution protocol (ARP) is used for **1 point**

- Finding the IP address from the DNS
- Finding the IP address of the default gateway
- Finding the IP address that corresponds to a MAC address
- Finding the MAC address that corresponds to an IP address

9) Consider different activities related to email: **1 point**

m1: Send an email from a mail client to a mail server

m2: Download an email from mailbox server to a mail client

m3: Checking email in a web browser

Which is the application level protocol used in each activity ?

- m1: HTTP m2: SMTP m3: POP
- m1: SMTP m2: FTP m3: HTTP
- m1: SMTP m2: POP m3: HTTP
- m1: POP m2: SMTP m3: IMAP

10) A process executes the code

1 point

```
fork();  
fork();  
fork();
```

The total number of child processes created is

- 3
- 4
- 7
- 8

11) Which of the following page replacement algorithms suffers from Belady's anomaly? **1 point**

- FIFO
- LRU
- Optimal Page Replacement
- Both LRU and FIFO

12) Suppose the numbers 7, 5, 1, 8, 3, 6, 0, 9, 4, 2 are inserted in that order into an initially empty binary search tree. The binary search tree uses the usual ordering on natural numbers. What is the in-order traversal sequence of the resultant tree ? **1 point**

- 7 5 1 0 3 2 4 6 8 9
- 0 2 4 3 1 6 5 9 8 7
- 0 1 2 3 4 5 6 7 8 9
- 9 8 6 4 2 3 0 1 5 7

You may submit any number of times before the due date. The final submission will be considered for grading.

Submit Answers

Quiz Assignment-I Solutions: Big Data Computing (Week-1)

Q.1 _____ is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.

- A. Hadoop Common
- B. Hadoop Distributed File System (HDFS)
- C. Hadoop YARN
- D. Hadoop MapReduce

Answer: C) Hadoop YARN

Explanation:

Hadoop Common: It contains libraries and utilities needed by other Hadoop modules.

Hadoop Distributed File System (HDFS): It is a distributed file system that stores data on a commodity machine. Providing very high aggregate bandwidth across the entire cluster.

Hadoop YARN: It is a resource management platform responsible for managing compute resources in the cluster and using them in order to schedule users and applications. YARN is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes

Hadoop MapReduce: It is a programming model that scales data across a lot of different processes.

Q. 2 Which of the following tool is designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases ?

- A. Pig
- B. Mahout
- C. Apache Sqoop
- D. Flume

Answer: C) Apache Sqoop

Explanation: Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases

Q. 3 _____ is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

- A. Flume
- B. Apache Sqoop
- C. Pig
- D. Mahout

Answer: A) Flume

Explanation: Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and very flexible architecture based on streaming data flows. It's quite robust and fail tolerant, and it's really tunable to enhance the reliability mechanisms, fail over, recovery, and all the other mechanisms that keep the cluster safe and reliable. It uses simple extensible data model that allows us to apply all kinds of online analytic applications.

Q. 4 _____ refers to the connectedness of big data.

- A. Value
- B. Veracity
- C. Velocity
- D. Valence

Answer: D) Valence

Explanation: Valence refers to the connectedness of big data. Such as in the form of graph networks

Q. 5 Consider the following statements:

Statement 1: Volatility refers to the data velocity relative to timescale of event being studied

Statement 2: Viscosity refers to the rate of data loss and stable lifetime of data

- A. Only statement 1 is true
- B. Only statement 2 is true
- C. Both statements are true
- D. Both statements are false

Answer: D) Both statements are false

Explanation: The correct statements are:

Statement 1: Viscosity refers to the data velocity relative to timescale of event being studied

Statement 2: Volatility refers to the rate of data loss and stable lifetime of data

Q. 6 _____ refers to the biases, noise and abnormality in data, trustworthiness of data.

- A. Value
- B. Veracity
- C. Velocity
- D. Volume

Answer: B) Veracity

Explanation: Veracity refers to the biases ,noise and abnormality in data, trustworthiness of data.

Q. 7 _____ 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. Mahout
- C. Flume
- D. Impala

Answer: D) Impala

Explanation: Cloudera, Impala was designed specifically at Cloudera, and it's a query engine that runs on top of the Apache Hadoop. The project was officially announced at the end of 2012, and became a publicly available, open source distribution. Impala 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.

Q. 8 True or False ?

NoSQL databases store unstructured data with no particular schema

- A. True
- B. False

Answer: A) True

Explanation: While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL (Not Only SQL) to handle unstructured data. NoSQL databases store unstructured data with no particular schema

Q. 9 _____ is a highly reliable distributed coordination kernel , which can be used for distributed locking, configuration management, leadership election, and work queues etc.

- A. Apache Sqoop
- B. Mahout
- C. ZooKeeper
- D. Flume

Answer: C) ZooKeeper

Explanation: ZooKeeper is a central store of key value using which distributed systems can coordinate. Since it needs to be able to handle the load, Zookeeper itself runs on many machines.

Q. 10 True or False ?

MapReduce is a programming model and an associated implementation for processing and generating large data sets.

- A. True
- B. False

Answer: A) True

Quiz Assignment-II Solutions: Big Data Computing (Week-2)

Q. 1 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

Answer: C) Both statements are true

Q. 2 _____ is the slave/worker node and holds the user data in the form of Data Blocks.

- A. NameNode
- B. Data block
- C. Replication
- D. DataNode

Answer: D) DataNode

Explanation: NameNode works as a master server that manages the file system namespace and basically regulates access to these files from clients, and it also keeps track of where the data is on the DataNodes and where the blocks are distributed essentially. On the other hand DataNode is the slave/worker node and holds the user data in the form of Data Blocks.

Q. 3 _____ works as a master server that manages the file system namespace and basically regulates access to these files from clients, and it also keeps track of where the data is on the Data Nodes and where the blocks are distributed essentially.

- A. Name Node
- B. Data block
- C. Replication
- D. Data Node

Answer: A) Name Node

Explanation: Name Node works as a master server that manages the file system namespace and basically regulates access to these files from clients, and it also keeps track of where the data is on the Data Nodes and where the blocks are distributed essentially. On the other

hand Data Node is the slave/worker node and holds the user data in the form of Data Blocks.

Q. 4 The number of maps in MapReduce is usually driven by the total size of _____

- A. Inputs
- B. Outputs
- C. Tasks
- D. None of the mentioned

Answer: A) Inputs

Explanation: Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key 'l' and passes them to the Reduce function.

Q. 5 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

Answer: B) False

Explanation: The task tracker will communicate the progress and report the results to the job tracker.

Q. 6 Point out the correct statement in context of YARN:

- A. YARN is highly scalable.
- B. YARN enhances a Hadoop compute cluster in many ways
- C. YARN extends the power of Hadoop to incumbent and new technologies found within the data center
- D. All of the mentioned

Answer: D) All of the mentioned

Q. 7 Consider the pseudo-code for MapReduce's WordCount example (not shown here). 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. map() and reduce()
- D. The code does not have to be changed

Answer: A) Only map()

Explanation: The map function takes a value and outputs key:value pairs.

For instance, if we define a map function that takes a string and outputs the length of the word as the key and the word itself as the value then

map(steve) would return 5:steve and

map(savannah) would return 8:savannah.

This allows us to run the map function against values in parallel.

So we have to only adapt the map() function of pseudo code.

Q. 8 The namenode knows that the datanode is active using a mechanism known as

- A. Heartbeats
- B. Datapulse
- C. h-signal
- D. Active-pulse

Answer: A) heartbeats

Explanation: In Hadoop Name node and data node do communicate using Heartbeat. Therefore Heartbeat is the signal that is sent by the datanode to the namenode after the regular interval to time to indicate its presence, i.e. to indicate that it is alive.

Q. 9 True or False ?

HDFS performs replication, although it results in data redundancy?

- A. True
- B. False

Answer: True

Explanation: Once the data is written in HDFS it is immediately replicated along the cluster, so that different copies of data will be stored on different data nodes. Normally the Replication factor is 3 as due to this the data does not remain over replicated nor it is less.

Q. 10 _____ 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

Answer: A) Map

Explanation: Maps are the individual tasks that transform input records into intermediate records and reduce processes and merges all intermediate values associated per key.

Quiz Assignment-III Solutions: Big Data Computing (Week-3)

Q. 1 In Spark, a _____ is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost.

- A. Spark Streaming
- B. FlatMap
- C. Driver
- D. Resilient Distributed Dataset (RDD)

Answer: D) Resilient Distributed Dataset (RDD)

Explanation: Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes. Formally, an RDD is a read-only, partitioned collection of records. RDDs can be created through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be operated on in parallel.

Q. 2 Given the following definition about the join transformation in Apache Spark:

def join[W](other: RDD[(K, W)]): RDD[(K, (V, W))]

Where join operation is used for joining two datasets. When it is called on datasets of type (K, V) and (K, W), it returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key.

Output the result of **joinrdd**, when the following code is run.

```
val rdd1 = sc.parallelize(Seq(("m",55),("m",56),("e",57),("e",58),("s",59),("s",54)))  
val rdd2 = sc.parallelize(Seq(("m",60),("m",65),("s",61),("s",62),("h",63),("h",64)))  
val joinrdd = rdd1.join(rdd2)  
joinrdd.collect
```

- A. Array[(String, (Int, Int))] = Array((m,(55,60)), (m,(55,65)), (m,(56,60)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (h,(63,64)), (s,(54,61)), (s,(54,62)))
- B. Array[(String, (Int, Int))] = Array((m,(55,60)), (m,(55,65)), (m,(56,60)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (e,(57,58)), (s,(54,61)), (s,(54,62)))
- C. Array[(String, (Int, Int))] = Array((m,(55,60)), (m,(55,65)), (m,(56,60)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (s,(54,61)), (s,(54,62)))
- D. None of the mentioned

Answer: C) `Array[(String, (Int, Int))]` = `Array((m,(55,60)), (m,(55,65)), (m,(56,60)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (s,(54,61)), (s,(54,62)))`

Explanation: `join()` is transformation which returns an RDD containing all pairs of elements with matching keys in this and other. Each pair of elements will be returned as a `(k, (v1, v2))` tuple, where `(k, v1)` is in this and `(k, v2)` is in other.

Q. 3 Consider the following statements in the context of Spark:

Statement 1: Spark improves efficiency through in-memory computing primitives and general computation graphs.

Statement 2: Spark improves usability through high-level APIs in Java, Scala, Python and also provides an interactive shell.

- A. Only statement 1 is true
- B. Only statement 2 is true
- C. Both statements are true
- D. Both statements are false

Answer: C) Both statements are true

Explanation: Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala and Python, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming. Spark comes with several sample programs. Spark provides an interactive shell – a powerful tool to analyze data interactively. It is available in either Scala or Python language. Spark improves efficiency through in memory computing primitives. In in-memory computation, the data is kept in random access memory (RAM) instead of some slow disk drives and is processed in parallel. Using this we can detect a pattern, analyze large data. This has become popular because it reduces the cost of memory. So, in-memory processing is economic for applications.

Q. 4 True or False ?

Resilient Distributed Datasets (RDDs) are fault-tolerant and immutable.

- A. True
- B. False

Answer: True

Explanation: Resilient Distributed Datasets (RDDs) are:

1. Immutable collections of objects spread across a cluster

2. Built through parallel transformations (map, filter, etc.)
3. Automatically rebuilt on failure
4. Controllable persistence (e.g. caching in RAM)

Q. 5 Which of the following is not a NoSQL database?

- A. HBase
- B. Cassandra
- C. SQL Server
- D. None of the mentioned

Answer: C) SQL Server

Explanation: NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.

Q. 6 True or False ?

Apache Spark potentially run batch-processing programs up to 100 times faster than Hadoop MapReduce in memory, or 10 times faster on disk.

- A. True
- B. False

Answer: True

Explanation: The biggest claim from Spark regarding speed is that it is able to "run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk." Spark could make this claim because it does the processing in the main memory of the worker nodes and prevents the unnecessary I/O operations with the disks. The other advantage Spark offers is the ability to chain the tasks even at an application programming level without writing onto the disks at all or minimizing the number of writes to the disks.

Q. 7 _____ leverages Spark Core fast scheduling capability to perform streaming analytics.

- A. MLlib
- B. Spark Streaming
- C. GraphX
- D. RDDs

Answer: B) Spark Streaming

Explanation: Spark Streaming ingests data in mini-batches and performs RDD transformations on those mini-batches of data.

Q. 8 _____ is a distributed graph processing framework on top of Spark.

- A. MLlib
- B. Spark streaming
- C. GraphX
- D. All of the mentioned

Answer: C) GraphX

Explanation: GraphX is Apache Spark's API for graphs and graph-parallel computation. It is a distributed graph processing framework on top of Spark.

Q. 9 Point out the incorrect statement in the context of Cassandra:

- A. It is a centralized key-value store
- B. It is originally designed at Facebook
- C. It is designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure
- D. It uses a ring-based DHT (Distributed Hash Table) but without finger tables or routing

Answer: A) It is a centralized key-value store

Explanation: Cassandra is a distributed key-value store.

Q. 10 Consider the following statements:

Statement 1: Scale out means grow your cluster capacity by replacing with more powerful machines.

Statement 2: Scale up means incrementally grow your cluster capacity by adding more COTS machines (Components Off the Shelf).

- A. Only statement 1 is true
- B. Only statement 2 is true
- C. Both statements are false
- D. Both statements are true

Answer: C) Both statements are false

Explanation: The correct statements are:

Scale up = grow your cluster capacity by replacing with more powerful machines

Scale out = incrementally grow your cluster capacity by adding more COTS machines (Components Off the Shelf)

Quiz Assignment-IV Solutions: Big Data Computing (Week-4)

Q. 1 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

Answer: B) P: Availability, Q: Consistency, R: Partition tolerance

Explanation:

CAP Theorem states following properties:

Consistency: All nodes see same data at any time, or reads return latest written value by any client.

Availability: The system allows operations all the time, and operations return quickly.

Partition-tolerance: The system continues to work in spite of network partitions.

Q. 2 Cassandra uses a protocol called _____ to discover location and state information about the other nodes participating in a Cassandra cluster.

- A. Key-value
- B. Memtable
- C. Heartbeat
- D. Gossip

Answer: D) Gossip

Explanation: Cassandra uses a protocol called gossip to discover location and state information about the other nodes participating in a Cassandra cluster. Gossip is a peer-to-peer communication protocol in which nodes periodically exchange state information about themselves and about other nodes they know about.

Q. 3 In Cassandra, _____ is used to specify data centers and the number of replicas to place within each data center. It attempts to place replicas on distinct racks to avoid the node failure and to ensure data availability.

- A. Simple strategy
- B. Quorum strategy
- C. Network topology strategy
- D. None of the mentioned

Answer: C) Network topology strategy

Explanation: Network topology strategy is used to specify data centers and the number of replicas to place within each data center. It attempts to place replicas on distinct racks to avoid the node failure and to ensure data availability. In network topology strategy, the two most common ways to configure multiple data center clusters are: Two replicas in each data center, and Three replicas in each data center.

Q. 4 True or False ?

A Snitch determines which data centers and racks nodes belong to. Snitches inform Cassandra about the network topology so that requests are routed efficiently and allows Cassandra to distribute replicas by grouping machines into data centers and racks.

- A. True
- B. False

Answer: True

Explanation: A Snitch determines which data centers and racks nodes belong to. Snitches inform Cassandra about the network topology so that requests are routed efficiently and allows Cassandra to distribute replicas by grouping machines into data centers and racks. Specifically, the replication strategy places the replicas based on the information provided by the new snitch. All nodes must return to the same rack and data center. Cassandra does its best not to have more than one replica on the same rack (which is not necessarily a physical location).

Q. 5 Consider the following statements:

Statement 1: In Cassandra, during a write operation, when hinted handoff is enabled and If any replica is down, the coordinator writes to all other replicas, and keeps the write locally until down replica comes back up.

Statement 2: In Cassandra, Ec2Snitch is important snitch for deployments and it is a simple snitch for Amazon EC2 deployments where all nodes are in a single region. In Ec2Snitch region name refers to data center and availability zone refers to rack in a cluster.

- A. Only Statement 1 is true
- B. Only Statement 2 is true
- C. Both Statements are true
- D. Both Statements are false

Answer: C) Both Statements are true

Explanation: Cassandra uses a protocol called gossip to discover location and state information about the other nodes participating in a Cassandra cluster. Gossip is a peer-to-peer communication protocol in which nodes periodically exchange state information about themselves and about other nodes they know about.

Q. 6 What is Eventual Consistency ?

- A. At any time, the system is linearizable
- B. If writes stop, all reads will return the same value after a while
- C. At any time, concurrent reads from any node return the same values
- D. If writes stop, a distributed system will become consistent

Answer: B) If writes stop, all reads will return the same value after a while

Explanation: Cassandra offers Eventual Consistency. It says that If writes to a key stop, all replicas of key will converge automatically.

Q. 7 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 false
- D. Both Statements are true

Answer: C) Both Statements are false

Explanation: The correct statements are:

Statement 1: When two processes are competing with each other causing data corruption, it is called Race Condition

Statement 2: When two processes are waiting for each other directly or indirectly, it is called deadlock.

Q. 8 ZooKeeper allows distributed processes to coordinate with each other through registers, known as _____

- A. znodes
- B. hnodes
- C. vnodes
- D. rnodes

Answer: A) znodes

Explanation: Every znode is identified by a path, with path elements separated by a slash.

Q. 9 In Zookeeper, when a _____ is triggered the client receives a packet saying that the znode has changed.

- A. Event
- B. Row
- C. Watch
- D. Value

Answer: C) Watch

Explanation: ZooKeeper supports the concept of watches. Clients can set a watch on a znodes.

Q. 10 Consider the Table temperature_details in Keyspace “day3” with schema as follows:

temperature_details(daynum, year,month,date,max_temp)

with **primary key(daynum,year,month,date)**

DayNum	Year	Month	Date	MaxTemp (°C)
1	1943	10	1	14.1
2	1943	10	2	16.4
541	1945	3	24	21.1
9970	1971	1	16	21.4
20174	1998	12	24	36.7
21223	2001	11	7	16
4317	1955	7	26	16.7

There exists same maximum temperature at different hours of the same day. Choose the correct CQL query to:

Alter table temperature_details to add a new column called “seasons” using map of type <varint, text> represented as <month, season>. Season can have the following values season={spring, summer, autumn, winter}.

Update table temperature_details where columns daynum, year, month, date contain the following values- 4317,1955,7,26 respectively.

Use the select statement to output the row after updation.

Note: A map relates one item to another with a key-value pair. For each key, only one value may exist, and duplicates cannot be stored. Both the key and the value are designated with a data type.

A)

```
cqlsh:day3> alter table temperature_details add hours1 set<varint>;
cqlsh:day3> update temperature_details set hours1={1,5,9,13,5,9} where daynum=4317;
cqlsh:day3> select * from temperature_details where daynum=4317;
```

B)

```
cqlsh:day3> alter table temperature_details add seasons map<varint,text>;
cqlsh:day3> update temperature_details set seasons = seasons + {7:'spring'} where
daynum=4317 and year=1955 and month=7 and date=26;
cqlsh:day3> select * from temperature_details where daynum=4317 and year=1955 and
month=7 and date=26;
```

C)

```
cqlsh:day3> alter table temperature_details add hours1 list<varint>;
cqlsh:day3> update temperature_details set hours1=[1,5,9,13,5,9] where daynum=4317 and
year=1955 and month=7 and date=26;
cqlsh:day3> select * from temperature_details where daynum=4317 and year=1955 and
month=7 and date=26;
```

D) cqlsh:day3> alter table temperature_details add seasons map<month, season>;
cqlsh:day3> update temperature_details set seasons = seasons + {7:'spring'} where
daynum=4317;
cqlsh:day3> select * from temperature_details where daynum=4317;

Answer: B)

```
cqlsh:day3> alter table temperature_details add seasons map<varint,text>;
cqlsh:day3> update temperature_details set seasons = seasons + {7:'spring'} where
daynum=4317 and year=1955 and month = 7 and date=26;
cqlsh:day3> select * from temperature_details where daynum=4317 and year=1955 and
month=7 and date=26;
```

Explanation:

The correct steps are:

a) Add column “seasons”

```
cqlsh:day3> alter table temperature_details add seasons map<varint,text>;
```

b) Update table

```
cqlsh:day3> update temperature_details set seasons = seasons + {7:'spring'} where
daynum=4317 and year=1955 and month = 7 and date=26;
```

c) Select query

```
cqlsh:day3> select * from temperature_details where daynum=4317 and year=1955 and
month=7 and date=26;
```

daynum	year	month	date	hours	hours1	max_temp	seasons
4317	1955	7	26	{1,5,9,13}	[1,5,9,13,5,9]	16.7	{7:'spring'}

Quiz Assignment-V Solutions: Big Data Computing (Week-5)

Q. 1 Columns in HBase are organized to_____

- A. Column group
- B. Column list
- C. Column base
- D. Column families

Answer: D) Column families

Explanation: An HBase table is made of column families which are the logical and physical grouping of columns. The columns in one family are stored separately from the columns in another family.

Q. 2 HBase is a distributed _____ database built on top of the Hadoop file system.

- A. Row-oriented
- B. Tuple-oriented
- C. Column-oriented
- D. None of the mentioned

Answer: C) Column-oriented

Explanation: A distributed column-oriented data store that can scale horizontally to 1,000s of commodity servers and petabytes of indexed storage.

Q. 3 A small chunk of data residing in one machine which is part of a cluster of machines holding one HBase table is known as_____

- A. Region
- B. Split
- C. Rowarea
- D. Tablearea

Answer : A) Region

Explanation: In Hbase, table Split into regions and served by region servers.

Q. 4 In HBase, _____is a combination of row, column family, column qualifier and contains a value and a timestamp.

- A. Cell
- B. Stores
- C. HMaster
- D. Region Server

Answer: A) Cell

Explanation: Data is stored in HBASE tables Cells and Cell is a combination of row, column family, column qualifier and contains a value and a timestamp.

Q. 5 HBase architecture has 3 main components:

- A. Client, Column family, Region Server
- B. Cell, Rowkey, Stores
- C. HMaster, Region Server, Zookeeper
- D. HMaster, Stores, Region Server

Answer: C) HMaster, Region Server, Zookeeper

Explanation: HBase architecture has 3 main components: HMaster, Region Server, Zookeeper.

1. HMaster: The implementation of Master Server in HBase is HMaster. It is a process in which regions are assigned to region server as well as DDL (create, delete table) operations. It monitors all Region Server instances present in the cluster.
2. Region Server: HBase Tables are divided horizontally by row key range into Regions. Regions are the basic building elements of HBase cluster that consists of the distribution of tables and are comprised of Column families. Region Server runs on HDFS DataNode which is present in Hadoop cluster.
3. Zookeeper: It is like a coordinator in HBase. It provides services like maintaining configuration information, naming, providing distributed synchronization, server failure notification etc. Clients communicate with region servers via zookeeper.

Q. 6 HBase stores data in _____

- A. As many filesystems as the number of region servers
- B. One filesystem per column family
- C. A single filesystem available to all region servers
- D. One filesystem per table.

Answer : C) A single filesystem available to all region servers

Q. 7 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

Answer: C) Broker

Explanation: A Kafka broker allows consumers to fetch messages by topic, partition and offset. Kafka broker can create a Kafka cluster by sharing information between each other directly or indirectly using Zookeeper. A Kafka cluster has exactly one broker that acts as the Controller.

Q. 8 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

Answer: C) Both statements are true

Q. 9 _____ is a central hub to transport and store event streams in real time.

- A. Kafka Core
- B. Kafka Connect
- C. Kafka Streams
- D. None of the mentioned

Answer: A) Kafka Core

Explanation: Kafka Core is a central hub to transport and store event streams in real time.

Q. 10 What are the parameters defined to specify window operation ?

- A. State size, window length
- B. State size, sliding interval
- C. Window length, sliding interval
- D. None of the mentioned

Answer: C) Window length, sliding interval

Explanation:

Following parameters are used to specify window operation:

- i) Window length: duration of the window
 - (ii) Sliding interval: interval at which the window operation is performed
- Both the parameters must be a multiple of the batch interval

Q. 11 Consider the following dataset Customers:

Name	Date	AmountSpent (In Rs.)
Alice	2020-05-01	50
Bob	2020-05-04	29
Bob	2020-05-01	25
Alice	2020-05-04	55
Alice	2020-05-03	45
Bob	2020-05-06	27

Using the **Customers** table answer the following using spark streaming fundamentals:

Using the following pseudo code, find the rank of each customer visiting the supermarket

```
val wSpec3 = Window.partitionBy("name").orderBy("date")
customers.withColumn( "rank", rank().over(wSpec3) ).show()
```

A)

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	1
Bob	2020-05-04	29	1
Bob	2020-05-01	25	1
Alice	2020-05-04	55	2
Alice	2020-05-03	45	2
Bob	2020-05-06	27	2

B)

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	3
Bob	2020-05-04	29	3
Bob	2020-05-01	25	2
Alice	2020-05-04	55	2
Alice	2020-05-03	45	2
Bob	2020-05-06	27	1

C)

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	Null
Alice	2020-05-03	45	50
Alice	2020-05-04	55	45
Bob	2020-05-01	25	Null
Bob	2020-05-04	29	25
Bob	2020-05-06	27	29

D)

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	1
Alice	2020-05-03	45	2
Alice	2020-05-04	55	3
Bob	2020-05-01	25	1
Bob	2020-05-04	29	2
Bob	2020-05-06	27	3

Answer: D)

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	1
Alice	2020-05-03	45	2
Alice	2020-05-04	55	3
Bob	2020-05-01	25	1
Bob	2020-05-04	29	2
Bob	2020-05-06	27	3

Explanation:

In this question, we want to know the order of a customer's visit (whether this is their first, second, or third visit).

// Create a window spec.

```
val wSpec3 = Window.partitionBy("name").orderBy("date")
```

In this window spec, the data is partitioned by customer. Each customer's data is ordered by date.

// The rank function returns what we want.

```
customers.withColumn( "rank", rank().over(wSpec3) ).show()
```

Name	Date	AmountSpent (In Rs.)	Rank
Alice	2020-05-01	50	1
Alice	2020-05-03	45	2
Alice	2020-05-04	55	3
Bob	2020-05-01	25	1
Bob	2020-05-04	29	2
Bob	2020-05-06	27	3

Q. 12 _____ is a Java library to process event streams live as they occur.

- A. Kafka Core
- B. Kafka Connect
- C. Kafka Streams
- D. None of the mentioned

Answer: C) Kafka Streams

Explanation: Kafka Streams is a Java library to process event streams live as they occur.

Quiz Assignment-VI Solutions: Big Data Computing (Week-6)

Q. 1 Which of the following is required by K-means clustering ?

- A. Defined distance metric
- B. Number of clusters
- C. Initial guess as to cluster centroids
- D. All of the mentioned

Answer: D) All of the mentioned

Explanation: K-means clustering follows partitioning approach.

Q. 2 Identify the correct statement in context of Regressive model of Machine Learning.

- A. Regressive model predicts a numeric value instead of category.
- B. Regressive model organizes similar item in your dataset into groups.
- C. Regressive model comes up with a set of rules to capture associations between items or events.
- D. None of the Mentioned

Answer: A) Regressive model predicts a numeric value instead of category.

Q. 3 Which of the following tasks can be best solved using Clustering ?

- A. Predicting the amount of rainfall based on various cues
- B. Training a robot to solve a maze
- C. Detecting fraudulent credit card transactions
- D. All of the mentioned

Answer: C) Detecting fraudulent credit card transactions

Explanation: Credit card transactions can be clustered into fraud transactions using unsupervised learning.

Q. 4 Identify the correct method for choosing the value of 'k' in k-means algorithm ?

- A. Dimensionality reduction
- B. Elbow method
- C. Both Dimensionality reduction and Elbow method
- D. Data partitioning

Answer: C) Both Dimensionality reduction and Elbow method

Q. 5 Identify the correct statement(s) in context of overfitting in decision trees:

Statement I: The idea of Pre-pruning is to stop tree induction before a fully grown tree is built, that perfectly fits the training data.

Statement II: The idea of Post-pruning is to grow a tree to its maximum size and then remove the nodes using a top-bottom approach.

- A. Only statement I is true
- B. Only statement II is true
- C. Both statements are true
- D. Both statements are false

Answer: A) Only statement I is true

Explanation: With pre-pruning, the idea is to stop tree induction before a fully grown tree is built that perfectly fits the training data.

In post-pruning, the tree is grown to its maximum size, then the tree is pruned by removing nodes using a bottom up approach.

Q. 6 Which of the following options is/are true for K-fold cross-validation ?

- 1. Increase in K will result in higher time required to cross validate the result.
 - 2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.
 - 3. If $K=N$, then it is called Leave one out cross validation, where N is the number of observations.
-
- A. 1 and 2
 - B. 2 and 3
 - C. 1 and 3
 - D. 1, 2 and 3

Answer: D) 1,2 and 3

Explanation: Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Q. 7 Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	Predicted:	
	NO	YES
Actual:		
NO	50	10
Actual:		
YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions ?

- 1. Accuracy is ~0.91
- 2. Misclassification rate is ~ 0.91
- 3. False positive rate is ~0.95
- 4. True positive rate is ~0.95

- A. 1 and 3
- B. 2 and 4
- C. 2 and 3
- D. 1 and 4

Answer: D) 1 and 4

Explanation:

The Accuracy (correct classification) is $(50+100)/165$ which is nearly equal to 0.91.

The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be $100/105 = 0.95$ also known as “Sensitivity” or “Recall”

Q. 8 Identify the correct statement(s) in context of machine learning approaches:

Statement I: In supervised approaches, the target that the model is predicting is unknown or unavailable. This means that you have unlabeled data.

Statement II: In unsupervised approaches the target, which is what the model is predicting, is provided. This is referred to as having labeled data because the target is labeled for every sample that you have in your data set.

- A. Only Statement I is true
- B. Only Statement II is true

- C. Both Statements are false
- D. Both Statements are true

Answer: C) Both Statements are false

Explanation: The correct statements are:

Statement I: In supervised approaches the target, which is what the model is predicting, is provided. This is referred to as having labeled data because the target is labeled for every sample that you have in your data set.

Statement II: In unsupervised approaches, the target that the model is predicting is unknown or unavailable. This means that you have unlabeled data.

Quiz Assignment-VII Solutions: Big Data Computing (Week-7)

Q. 1 Suppose you are using a bagging based algorithm say a Random Forest in model building. Which of the following can be true?

- 1 Number of tree should be as large as possible
 - 2 You will have interpretability after using Random Forest
-
- A. Only 1
 - B. Only 2
 - C. Both 1 and 2
 - D. None of these

Answer: A) Only 1

Explanation: Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

Q. 2 To apply bagging to regression trees which of the following is/are true in such case?

1. We build the N regression with N bootstrap sample
 2. We take the average the of N regression tree
 3. Each tree has a high variance with low bias
-
- A. 1 and 2
 - B. 2 and 3
 - C. 1 and 3
 - D. 1,2 and 3

Answer: D) 1,2 and 3

Explanation: All of the options are correct and self explanatory

Q. 3 In which of the following scenario a gain ratio is preferred over Information Gain?

- A. When a categorical variable has very small number of category
- B. Number of categories is the not the reason
- C. When a categorical variable has very large number of category
- D. None of the mentioned

Answer: C) When a categorical variable has very large number of category

Explanation: When high cardinality problems, gain ratio is preferred over Information Gain technique.

Q. 4 Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?

1. Both methods can be used for classification task
2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
4. Both methods can be used for regression task
 - A. 1 and 2
 - B. 2 and 3
 - C. 2 and 4
 - D. 1 and 4

Answer: D) 1 and 4

Explanation: Both algorithms are design for classification as well as regression task.

Q. 5 Given an *attribute table* shown below, which stores the basic information of attribute *a*, including the row identifier of instance *row_id* , values of attribute *values* (*a*) and class labels of instances *c*.

Attribute Table			
Outlook	Humidity	Wind	Play
Sunny	High	Weak	No
Sunny	High	Strong	No
Overcast	High	Weak	Yes
Rain	High	Weak	Yes
Rain	Normal	Weak	Yes
Rain	Normal	Strong	No
Overcast	Normal	Strong	Yes
Sunny	High	Weak	No
Sunny	Normal	Weak	Yes
Rain	Normal	Weak	Yes
Sunny	Normal	Strong	Yes

Overcast	High	Strong	Yes
Overcast	Normal	Weak	Yes
Rain	High	Strong	No

Which of the following attribute will first provide the pure subset ?

- A. Humidity
- B. Wind
- C. Outlook
- D. None of the mentioned

Answer: C) Outlook

Explanation: To measure the pureness or uncertainty of a subset, we need to provide a quantitative measure so that the Decision Tree algorithm can be objective when choosing the best attribute and condition to split on. There are different ways to measure the uncertainty in a set of values, but for the purposes of this example, we will use Entropy (represented by "H").

$$H(X) = - \sum_{i=1}^n p_i \log p_i$$

Where X is the resulting split, n is the number of different target values in the subset, and p_i is the proportion of the i^{th} target value in the subset.

For example, the entropy will be the following. The log is base 2.

$$\text{Entropy (Sunny)} = -2/5 * \log(2/5) - 3/5 \log(3/5) = 0.159 + 0.133 = 0.292 \text{ (Impure subset)}$$

$$\text{Entropy (Overcast)} = -4/4 * \log 1 = 0 \text{ (Pure subset)}$$

$$\text{Entropy (Rain)} = -3/5 * \log(3/5) - 2/5 \log(2/5) = 0.292 \text{ (Impure subset)}$$

$$\text{Entropy (High)} = -3/7 * \log(3/7) - 4/7 \log(4/7) = 0.158 + 0.138 = 0.296 \text{ (Impure subset)}$$

$$\text{Entropy (Normal)} = -6/7 * \log(6/7) - 1/7 \log(1/7) = 0.057 + 0.121 = 0.177 \text{ (Impure subset)}$$

$$\text{Entropy (Weak)} = -6/8 * \log(6/8) - 2/8 \log(2/8) = 0.093 + 0.150 = 0.243 \text{ (Impure subset)}$$

$$\text{Entropy (Strong)} = -3/6 * \log(3/6) - 3/6 \log(3/6) = 0.15 + 0.15 = 0.30 \text{ (Impure subset)}$$

Q. 6 True or False ?

Bagging provides an averaging over a set of possible datasets, removing noisy and non-stable parts of models.

- A. True
- B. False

Answer: A) True

Q. 7 Hundreds of trees can be aggregated to form a Random forest model. Which of the following is true about any individual tree in Random Forest?

- 1. Individual tree is built on a subset of the features
 - 2. Individual tree is built on all the features
 - 3. Individual tree is built on a subset of observations
 - 4. Individual tree is built on full set of observations
-
- A. 1 and 3
 - B. 1 and 4
 - C. 2 and 3
 - D. 2 and 4

Answer: A) 1 and 3

Explanation: Random forest is based on bagging concept, that consider fraction of sample and fraction of feature for building the individual trees.

Q. 8 Boosting any algorithm takes into consideration the weak learners. Which of the following is the main reason behind using weak learners?

Reason I-To prevent overfitting

Reason II- To prevent underfitting

- A. Reason I
- B. Reason II
- C. Both Reason I and Reason II
- D. None of the Reasons

Answer: A) Reason I

Explanation: To prevent overfitting, since the complexity of the overall learner increases at each step. Starting with weak learners implies the final classifier will be less likely to overfit.

Quiz Assignment-VIII Solutions: Big Data Computing (Week-8)

Q. 1 Which of the following are provided by spark API for graph parallel computations:

- i. joinVertices
 - ii. subgraph
 - iii. aggregateMessages
- A. Only (i)
B. Only (i) and (ii)
C. Only (ii) and (iii)
D. All of the mentioned

Answer: D) All of the mentioned

Q. 2 Which of the following statement(s) is/are true in the context of Apache Spark GraphX operators ?

S1: Property operators modify the vertex or edge properties using a user defined map function and produces a new graph.

S2: Structural operators operate on the structure of an input graph and produces a new graph.

S3: Join operators add data to graphs and produces a new graphs.

- A. Only S1 is true
- B. Only S2 is true
- C. Only S3 is true
- D. All of the mentioned

Answer: D) All of the mentioned

Q. 3 True or False ?

The outerJoinVertices() operator joins the input RDD data with vertices and returns a new graph. The vertex properties are obtained by applying the user defined map() function to the all vertices, and includes ones that are not present in the input RDD.

- A. True
- B. False

Answer: A) True

Q. 4 Which of the following statements are true ?

S1: Apache Spark GraphX provides the following property operators - mapVertices(), mapEdges(), mapTriplets()

S2: The RDDs in Spark, depend on one or more other RDDs. The representation of dependencies in between RDDs is known as the lineage graph. Lineage graph information is used to compute each RDD on demand, so that whenever a part of persistent RDD is lost, the data that is lost can be recovered using the lineage graph information.

- A. Only S1 is true
- B. Only S2 is true
- C. Both S1 and S2 are true
- D. None of the mentioned

Answer: C) Both S1 and S2 are true

Q. 5 GraphX provides an API for expressing graph computation that can model the _____ abstraction.

- A. GaAdt
- B. Pregel
- C. Spark Core
- D. None of the mentioned

Answer: B) Pregel

Q. 6 Match the following:

- | | |
|--------------------------|-----------------------|
| A. Dataflow Systems | i. Vertex Programs |
| B. Graph Systems | ii. Parameter Servers |
| C. Shared Memory Systems | iii. Guinea Pig |

- A. A:ii, B: i, C: iii
- B. A:iii, B: i, C: ii
- C. A:ii, B: iii, C: i
- D. A:iii, B: ii, C: i

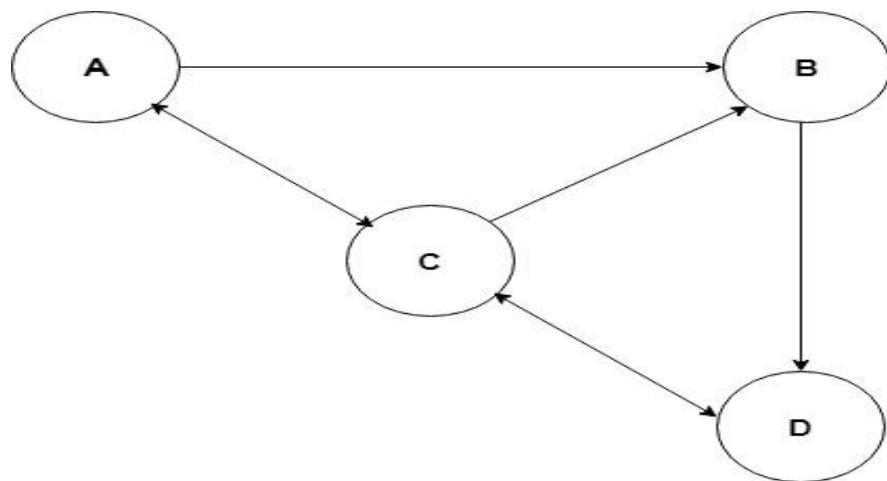
Answer: B) A:iii, B: i, C: ii

Q. 7 Which of the following statement(s) is/are true in context of Parameter Servers.

- S1: A machine learning framework
S2: Distributes a model over multiple machines
S3: It offers two operations: (i) Pull for query parts of the model (ii) Push for update parts of the model.
- A. Only S1 is true
B. Only S2 is true
C. Only S3 is true
D. All of the mentioned

Answer: D) All of the mentioned

Q. 8



What is the PageRank score of vertex **B** after the second iteration? (Without damping factor)

Hint:- The basic PageRank formula is:

$$PR_{t+1}(u) = \sum PR_t(v) / C(v)$$

Where, $PR_{t+1}(u)$: page rank of node u under consideration
 $PR_t(v)$: previous page rank of node ' v ' pointing to node ' u '
 $C(v)$: outgoing degree of vertex ' v '

- A. 1/6
B. 1.5/12
C. 2.5/12
D. 1/3

Answer: A) 1/6

Explanation: The Page Rank score of all vertex is calculated as follows:

	Iteration 0	Iteration 1	Iteration 2	Page Rank
A	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1.5}{12}$	1
B	$\frac{1}{4}$	$\frac{2.5}{12}$	$\frac{2}{12}$	2
C	$\frac{1}{4}$	$\frac{4.5}{12}$	$\frac{4.5}{12}$	4
D	$\frac{1}{4}$	$\frac{4}{12}$	$\frac{4}{12}$	3
