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

Q. 1 True or False ?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

- True
- False

Answer: True

Q2. What does "Velocity" in Big Data mean?

- Speed of input data generation
- Speed of individual machine processors
- Speed of storing and processing data
- Speed of ONLY storing data

Answer: Speed of storing and processing data

Explanation: Velocity is the speed at which data is processed. This includes input such as processing of social media posts and output such as the processing required to produce a report or execute a process.

Q. 3 _______________ refers to the accuracy and correctness of the data relative to a

particular use.

- Value
- Veracity
- Velocity
- Validity

Answer: Validity

Explanation: Validity refers to the accuracy and correctness of the data relative to a particular use.

Q. 4 Consider the following statements:

Statement 1: Viscosity refers to the connectedness of big data.

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

- Only statement 1 is true
- Only statement 2 is true
- Both statements are true
- Both statements are false

Answer: Only statement 2 is true

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

- HDFS
- YARN
- Map Reduce
- PIG

Answer: Map Reduce

Explanation: Map Reduce is a programming model and an associated implementation for processing and generating large data sets.

Q. 6 _______________is an open source software framework for big data. It has two basic parts: HDFS and Map Reduce.

- Spark
- HBASE
- HIVE
- Apache Hadoop

Answer: Apache Hadoop

Explanation: Apache Hadoop is an open source software framework for big data

It has two basic parts:

a) Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster

b) MapReduce: Programming model that simplifies parallel programming

Q.7 The fundamental idea of __________________ is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. The idea is to have a global Resource Manager (RM) and per-application Application Master (AM). An application is either a single job or a DAG of jobs.

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop YARN
- Hadoop MapReduce

Answer: 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. 8 ____________is a highly reliable distributed coordination kernel , which can be used for distributed locking, configuration management, leadership election, and work queues etc.

- Apache Sqoop
- Mahout
- Flume
- ZooKeeper

Answer: 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. 9 ______________ is an open source stream processing software platform developed by the Apache Software Foundation written in Scala and Java.

- Hive
- Cassandra
- Apache Kafka
- RDDs

Answer: Apache Kafka

Explanation: Apache Kafka is an open source stream processing software platform developed by the Apache Software Foundation written in Scala and Java.

Q. 10 True or False ?

NoSQL databases are non-tabular databases and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph. They provide flexible schemas and scale easily with large amounts of data and high user loads.

- True
- False

Answer: 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

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

Q. 1 _______________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. Data Node
B. Name Node
C. Data block
D. Replication

Answer: B. 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. 2 - When a client contacts the name node for accessing a file, the name node responds with

A. Size of the file requested.
B. Block ID of the file requested.
C. Block ID and hostname of any one of the data nodes containing that block.
D. Block ID and hostname of all the data nodes containing that block.

Answer: D. Block ID and hostname of all the data nodes containing that block.

Explanation: A name node is 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.

Q. 3 The namenode knows that the datanode is active using a mechanism known as

A. datapulse
B. h-signal
C. heartbeats
D. Active-pulse

Answer: C. 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. 4 For reading/writing data to/from HDFS, clients first connect to ______________

A. NameNode
B. Checkpoint Node
C. DataNode
D. None of the mentioned

Answer: A. NameNode

Explanation: To read/write a file in HDFS, a client needs to interact with master i.e. namenode (master).

Q. 5 True or False ?

HDFS performs replication, although it results in data redundancy?

A. True
B. False

Answer: A) 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. 6 Consider the following statements:

Statement 1: Task Tracker is hosted inside the master and it receives the job execution request from the client.

Statement 2: Job 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: D. Both statements are false

Explanation: The correct statements are:

The Job Tracker is hosted inside the master and it receives the job execution request from the client.
Task tracker is the MapReduce component on the slave machine as there are multiple slave machines.

Q. 7 Consider the following statements:

Statement 1: MapReduce is a programming model and an associated implementation for processing and generating large data sets.

Statement 2: Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key.

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. 8 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

Answer: D. All of the mentioned

Q. 9 Apache Hadoop YARN stands for:

A. Yet Another Reserve Negotiator
B. Yet Another Resource Negotiator
C. Yet Another Resource Network
D. Yet Another Resource Manager

Answer: B. Yet Another Resource Negotiator

Q 10. Consider the pseudo-code for MapReduce's WordCount example (not shown here). Let's now assume that you want to determine the average length of all the words in a text file. 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: C. map() and reduce()

Explanation:

The problem statement is:

Assume a given file name: sample.txt

File contents:

This is a sample file

Fit for nothing

Length of words

This = 4, is =2, a=1, sample=6, file=4, fit=3, for=3, nothing=7

Total length of all words=4+2+1+6+4+3+3+7=30

Total number of words=8

Average length of all the words= total length of all the words/total number of words=30/8=3.75

Expected output= 3.75

Programming the Mapper

Mapper is programmed to do the following:

Step-1: Ignore the key from the recorder

Step 2: Split the words in the value (the full line)

This is a sample file

[this] [is] [a] [sample][file] (line is split)

Step 3: Output the number 1 as key each word as value. The sample output of mapper would look like

KEY: 1, VALUE: this

Step 4: Repeat the above steps for all the words in the line

Output of the Mapper after processing the entire file

KEY	VALUE
1	this
1	is
1	a
1	sample
1	file
1	fit
1	for
1	nothing

All the keys are 1 here. So the output would look like…

KEY	VALUE
1	this is a sample file fit for nothing

The reducer would just contain 1 key and a list of values.

All the words in the file would indeed be in the list of values passed on to reducer.

And finally compute the average length of each word since we know the total length of all the words and also the total count of words.

Hence, we need both map() and reduce() to adapt.

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

Q. 1 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. 2 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. 3 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. Resilient Distributed Dataset (RDD)
D. Driver

Answer: C) 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.

Q. 4 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)), (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)), (e,(57,58)), (s,(54,61)), (s,(54,62)))

D. None of the mentioned

Answer: B) 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. 5 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: A) 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. 6 **______________** leverages Spark Core fast scheduling capability to perform streaming analytics.

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

Answer: D) Spark Streaming

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

Q. 7 **____________________** is a distributed graph processing framework on top of Spark.

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

Answer: A) 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. 8 Which of the following are the simplest NoSQL databases ?

A. Wide-column
B. Key-value
C. Document
D. All of the mentioned

Answer: B) Key-value

Explanation: Every single item in the database is stored as an attribute name (or “key”), together with its value in Key-value stores.

Q. 9 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)

Q. 10 Point out the incorrect statement in the context of Cassandra:

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

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

Explanation: Cassandra is a distributed key-value store.

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

Q. 1 In Cassandra, ________________ replication strategy treats the entire cluster as a single data center. It is suitable for a single data center and one rack. This is also called as Rack-Unaware Strategy.

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

Answer: A) Simple strategy

Explanation: Simple strategy treats the entire cluster as a single data center. It is suitable for a single data center and one rack. This is also called as Rack-Unaware Strategy. Simple Strategy uses the partitioner of which there are two kinds: Random Partitioner and Byte Ordered Partitioner.

Q. 2 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. Network topology strategy
C. Quorum strategy
D. None of the mentioned

Answer: B) 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. 3 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. 4 Identify the correct choices for the given scenarios:

P: All nodes see same data at any time, or reads return latest written value by any client
Q: The system allows operations all the time, and operations return quickly
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: A) P: Consistency, Q: Availability, 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. 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

Q. 6 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. Gossip
D. Heartbeat

Answer: C) 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. 7 What is Eventual Consistency ?

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

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

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

Q. 8 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. 9 Which of the following is incorrect statement ?

A. ZooKeeper is a distributed co-ordination service to manage large set of hosts.
B. ZooKeeper allows developers to focus on core application logic without worrying about the distributed nature of the application.
C. ZooKeeper solves this issue with its simple architecture and API.
D. The ZooKeeper framework was originally built at "Google" for accessing their applications in an easy and robust manner.

Answer: D) The ZooKeeper framework was originally built at "Google" for accessing their applications in an easy and robust manner

Explanation: The ZooKeeper framework was originally built at "Yahoo!" for accessing their applications in an easy and robust manner

Q. 10 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. 11 ZooKeeper itself is intended to be replicated over a sets of hosts called ____________

A. Chunks
B. Ensemble
C. Subdomains
D. None of the mentioned

Answer: B) Ensemble

Explanation: ZooKeeper run on a cluster of machines called an ensemble. As long as a majority of the servers are available, the ZooKeeper service will be available.

Q. 12 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 True or False ?

Apache HBase is a column-oriented, NoSQL database designed to operate on top of the Hadoop distributed file system (HDFS).

A. True
B. False

Answer: A) True

Explanation: Apache HBase is a column-oriented NoSQL database that runs on top of the Hadoop Distributed File System, a main component of Apache Hadoop

Q. 2 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. Rowarea
B. Tablearea
C. Region
D. Split

Answer : C) Region

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

Q. 3 In HBase, what is the number of MemStore per column family ?

A. 1

B. 2

C. 3

D. Equal to as many columns in the column family

Answer : A) 1

Explanation: There is only one Memstore per column family.

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

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

Answer: D) 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. HMaster, Region Server, Zookeeper
C. Cell, Rowkey, Stores
D. HMaster, Stores, Region Server

Answer: B) 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 monitor 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 True or False ?

Kafka is a high performance, real time messaging system. It is an open source tool and is a part of Apache projects.

A. True
B. False

Answer: True

Explanation: Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications.

Q. 7 Kafka maintains feeds of messages in categories called____________________

A. Chunks
B. Domains
C. Messages
D. Topics

Answer: D) Topics

Explanation: A topic is a category or feed name to which messages are published. For each topic, the Kafka cluster maintains a partitioned log

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 What exactly Kafka key capabilities?

A. Publish and subscribe to streams of records, similar to a message queue or enterprise messaging system
B. Store streams of records in a fault-tolerant durable way
C. Process streams of records as they occur
D. All of the mentioned

Answer: D) All of the mentioned

Q. 10 __________________is a framework to import event streams from other source data systems into Kafka and export event streams from Kafka to destination data systems.

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

Answer: B) Kafka Connect

Explanation:
Kafka connect is a framework to import event streams from other source data systems into Kafka and export event streams from Kafka to destination data systems.

Q. 11 ________________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. 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 tasks can be best solved using Clustering ?

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

Answer: B) Detecting fraudulent credit card transactions

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

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.

Explanation: When your model has to predict a numeric value instead of a category, then the task becomes a regression problem. An example of regression is to predict the price of a stock. The stock price is a numeric value, not a category. So this is a regression task instead of a classification task.

Q. 3 ___________ refers to a model that can neither model the training data nor generalize to new data.

A. Good fitting
B. Overfitting
C. Underfitting
D. All of the mentioned

Answer: C) Underfitting

Explanation: An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data. Usually, a model that is underfit will have high training and high testing error.

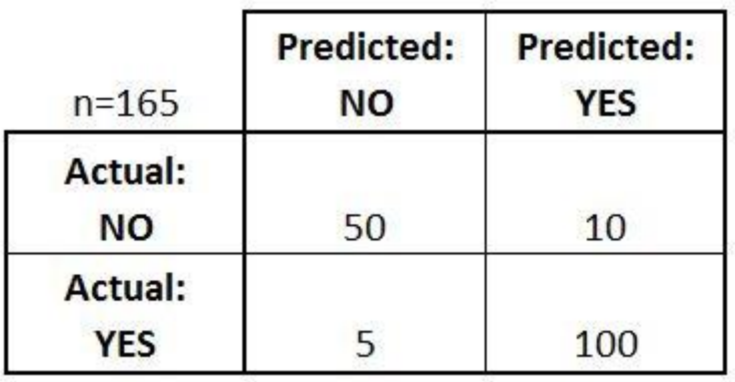
Q. 4 Which of the following is required by K-means clustering ?

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. 5 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	Predicted: 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. 1 and 4
C. 2 and 4
D. 2 and 3

Answer: B) 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. 6 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. 7 True or False ?

If your model has very low training error but high generalization error, then it is overfitting.

A. True
B. False

Answer: A) True

Explanation: A related concept to generalization is overfitting. If your model has very low training error but high generalization error, then it is overfitting. This means that the model has learned to model the noise in the training data, instead of learning the underlying structure of the data.

Q. 8 Identify the correct statement(s) in context of overfitting in decision trees:

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

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

A. Only Statement I is true
B. Only Statement II is true
C. Both Statements are true
D. Both Statements are false

Answer: B) Only Statement II is true

Explanation:

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

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

Q. 9 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. 10 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 True or False ?

The bootstrap sampling method is a resampling method that uses random sampling with replacement.

A. True
B. False

Answer: A) True

Explanation: The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. It can be used to estimate summary statistics such as the mean or standard deviation. It is used in applied machine learning to estimate the skill of machine learning models when making predictions on data not included in the training data.

Q. 2 True or False ?

Statement 1: Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.

Statement 2: Bagging provides an averaging over a set of possible datasets, removing noisy and non-stable parts of models.

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: Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

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

Answer: B) 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 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. Outlook
C. Wind
D. None of the mentioned

Answer: B) 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 $\mathbf{p}_i$ is the proportion of the $\boldsymbol{i^{th}}$ target value in the subset.
For example, the entropy will be the following. The log is base 2.

Entropy (Sunny) = -2/5 * log(2/5) – 3/5 log(3/5) =0.159+0.133= 0.292 (Impure subset)

Entropy (Overcast) = -4/4 * log 1 = 0 (Pure subset)

Entropy (Rain) = -3/5 * log(3/5) – 2/5 log(2/5) = 0.292 (Impure subset)

Entropy (High) = -3/7 * log(3/7) – 4/7 log(4/7) = 0.158+0.138=0.296 (Impure subset)

Entropy (Normal) = -6/7 * log(6/7) – 1/7 log(1/7) = 0.057+0.121=0.177 (Impure subset)

Entropy (Weak) = -6/8 * log(6/8) – 2/8 log(2/8) = 0.093+0.150=0.243 (Impure subset)

Entropy (Strong) = -3/6 * log(3/6) – 3/6 log(3/6) = 0.15+0.15=0.30 (Impure subset)

Q. 5 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 faction of sample and faction of feature for building the individual trees.

Q. 6 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. 1 and 4
   D. 2 and 4

Answer: C) 1 and 4

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

Q. 7 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 the Reasons
   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.

Q. 8 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

---

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

Q. 1 Which of the following statement(s) is/are true in the context of Apache Spark GraphX operators ?

S1: Structural operators operate on the structure of an input graph and produces a new graph.

S2: Property operators modify the vertex or edge properties using a user defined map function and produces a new graph.

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

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

Answer: D) All of the mentioned

Q. 2 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

Explanation: GraphX implements a Pregel-like bulk-synchronous message-passing API. Unlike the original Pregel API, the GraphX Pregel API factors the sendMessage computation over edges, enables the message sending computation to read both vertex attributes, and constrains messages to the graph structure. These changes allow for substantially more efficient distributed execution while also exposing greater flexibility for graph-based computation.

Q. 3 Match the following:

A. Dataflow Systems	i. Vertex Programs
B. Graph Systems	ii. Parameter Servers
C. Shared Memory Systems	iii. GuineaPig

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

Explanation:
First, dataflow systems such as Hadoop and spark. Data is independent records and push through a processing pipeline.
Graph systems like Graphlab, etc. Problem is modeled as a graph, each node communicates with its neighbors.
Distributed shared memory systems like Bosen, etc. Model is globally accessible and changed by external workers.

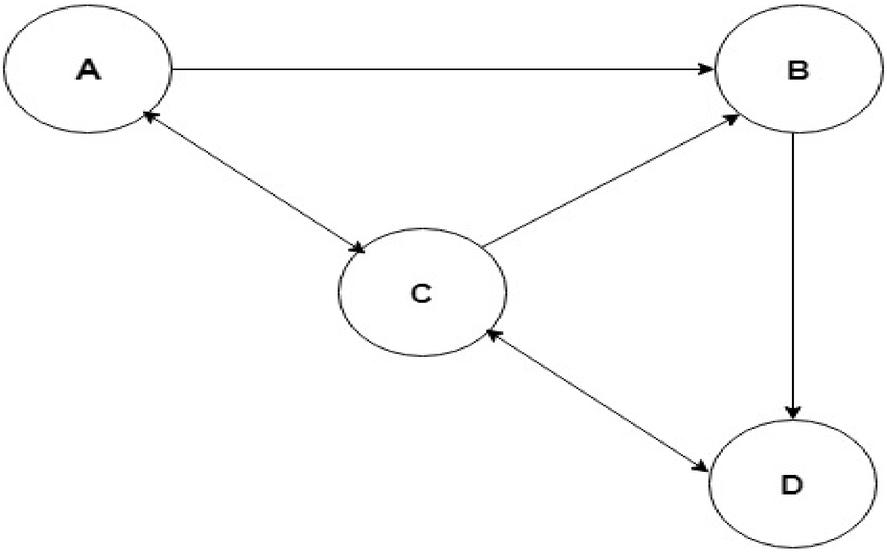
Each of these types of systems offer a different abstraction as well such as:

Dataflow Systems → PIG, GuineaPig

Graph Systems → Vertex-Programs

Shared Memory Systems → Parameter Servers

Q. 4 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	1/4	1/12	1.5/12	1
B	1/4	2.5/12	2/12	2
C	1/4	4.5/12	4.5/12	4
D	1/4	4/12	4/12	3

Q. 5 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 statement S1 is true
B. Only statement S2 is true
C. Only statement S3 is true
D. All of the mentioned

Answer: D) All of the mentioned

Explanation: Parameter Server is a machine learning framework. It distributes a model over multiple machines. It offers two operations:

1. Pull: query parts of the model
2. Push: update parts of the model

Q. 6 Identify the correct statement for Stale synchronous process (SSP):

Statement 1: SSP interpolates between BSP (Bulk synchronous parallel) and Asynchronous and subsumes both.

Statement 2: SSP allows usually workers to run at own pace

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: Following are the features for Stale Synchronous Parallel (SSP):

1. Interpolate between BSP and Async and subsumes both
2. Allow workers to usually run at own pace
3. Fastest/slowest threads not allowed to drift >s clocks apart
4. Efficiently implemented: Cache parameters

Q. 7 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. 8 Which of the following statement(s) is/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