Home / My courses / WQD7007 / Mid term test / Mid term test

Started on	Tuesday, 15 December 2020, 6:15 PM
State	Finished
Completed on	Tuesday, 15 December 2020, 7:42 PM
Time taken	1 hour 27 mins

Question **1**

Correct

Marked out of 1.00

Select the statements that are false:

i. Hadoop distributed file system (HDFS) is designed to run across commodity hardware.

ii. NameNode actively monitors the number of replicas of a block to make sure the number of replicas is adequate.

iii. NameNode send heartbeats to DataNodes to make sure the DataNodes are functioning.

iv. Client reads data from NameNode after DataNode send the necessary information to NameNode.

Select one:

○ a. ii and iv

○ b. ii and iii

● c. iii and iv ✓

○ d. i and ii

Question **2**

Not answered

Marked out of 1.00

Select the Hadoop tools that is used for data access:

i. YARN

ii. Hive

iii. Kafka

iv. Ambari

Select one:

○ a. ii and iv

○ b. i and ii

○ c. ii and iii

○ d. iii and iv

Question **3**

Complete

Marked out of 5.00

Suggest 3 bad identifier methods with appropriate examples. Explain why the suggested identifier methods are considered bad/poor to be used in identifier system.

i) Using a non unique identifier such as a person's name. There could be many people with the same name, for example they are millions of Chinese with "Lee" as their surname. This would violate the uniqueness in big data resources as one identifier might no longer point to an unique data object.

ii) Embedding extra information into the identifier, especially personal identifiable information. Under certain circumstances, one might utilize the information embedded to filter the dataset and subsequently identify the private records. For example, the IC number system of Malaysia embeds date of birth, gender and area of birth information.

iii) Assigning an identifier that is not properly scoped or prefixed for a resource that is expected to be centralized. Although the identifier works well initially, problems might arise when the big data manager attempts to merge multiple resources. For example, if multiple departments within an organization use the same running number for their inventories, there will be replicates when merging the dataset.

Question **4**

Correct

Marked out of 1.00

The following is true about classification, except:

Select one:

○ a. The members of classes may be highly similar to one another, but their similarities result from their membership in the same class, and not the other way around.

● b. Every instance belongs to more than one class ✓

○ c. In a hierarchical classification, each subclass may have no more than one parent class.

○ d. The classes of the hierarchy have a set of properties or rules that extend to every member of the class and to all of the subclasses of the class.

Question **5**

Complete

Marked out of 5.00

Explain what is a “stop word”. Then, give 3 examples of stop word, and explain why these words are suitable to be categorized as stop words.

"Stop words" are common terms that usually demarcate the uncommon terms, in which the uncommon terms are one or more words that carry the important concept. During term extraction, it is important to recognize stop words and filter them out, in order to get the uninterrupted sequences of uncommon words. Depending on the domain of interest for the big data resource, the list of stop words that are appropriate may be different.

For example, if we don’t filter out stop words, when we are building an index using term extraction algorithm, the resulting index will be full of words like “The”, “An”, “He”, which carry no important information.

Examples:

i) “The” - This is an obvious term that should be a stop word as it appears in most English sentences.

ii) “About” - Similar to the first example, this is a common English preposition that should be filtered out.

iii) “Disease” - For example, if we are extracting terms from a collection of medical articles about disease, this word should be the stop words as it would appear in most articles, although it might not be considered as one in another context.

Question **6**

Not answered

Marked out of 5.00

Assume that you are a data manager in an investment company, where you are required to analyze the potential changes in different investment sectors, and suggest optimum investment plan to the customer. You have a workstation configured to store different source of useful information on different investment sectors that capable of parallel processing, and you have administrator access to 10 employee's PC in the office, with no parallel processing capability.

In your opinion, determine whether you will use parallel computing or distributed computing to effectively solve your business problem. Justify your answer.

Question **7**

Not answered

Marked out of 1.00

To extract information from text, it is necessary to design structure for it. This may involve:

i. translating the text to a preferred language

ii. extracting and normalizing the conceptual terms contained in the sentences

iii. extracting and standardizing data values from the text

iv. assigning data values to specific classes of data belonging to a classification system

Select one:

○ a. ii, iii, and iv

○ b. i, ii, and iii

○ c. i, ii, iii and iv

○ d. i, iii, and iv

Question **8**

Incorrect

Marked out of 1.00

The following is false about Hadoop Distributed File System (HDFS), except:

Select one:

● a. NameNode will send heartbeat to DataNode every few seconds to make sure the DataNode is still available✗

○ b. There is only one NameNode in HDFS.

○ c. Multiple replicas are stored in the same DataNode.

○ d. Client reads data from NameNode after DataNode send the necessary information to NameNode.

Question **9**

Correct

Marked out of 1.00

Below are the disadvantages to use traditional file access compared to modern electronic database systems, except:

Select one:

● a. Complexity✓

○ b. Data manipulation

○ c. Accessibility

○ d. data order

Question **10**

Correct

Marked out of 1.00

Which of the following is correct about Variety as one of the V's in big data characteristics?

Select one:

● a. Different forms of data✓

○ b. Meaning of data in different environment.

○ c. Accurateness of the data.

○ d. Analysis of streaming data.

Question **11**

Not answered

Marked out of 1.00

Which of the following is wrong about small data?

Select one:

○ a. Contain structured data only.

○ b. Small data projects are easily replicable.

○ c. Usually designed to achieve a specific goal.

○ d. Multiple format can be used to represent the data.

Question **12**

Correct

Marked out of 1.00

Each of the following are true in explaining the features of an identifier system, except:

Select one:

● a. Uniqueness → Each identifier is assigned to a unique object, and to no other object. ✓

○ b. Security → A Big Data resource with an identifier system can be irreversibly corrupted if the identifiers are modified.

○ c. Completeness → Every unique object in the Big Data resource must be assigned an identifier.

○ d. Aggregation → The Big Data resource must have a mechanism to aggregate all of the data that is properly associated with the identifier.

Question **13**

Not answered

Marked out of 1.00

The following statements are true, except:

Select one:

○ a. Data importance principle is used to decide which data shall be stored and which data shall be discarded.

○ b. In big data architecture, professional tools are needed to kept the sensitive information safe.

○ c. Data representation helps to build more meaningful structure for computer analysis.

○ d. Redundancy level is kept high to reduce the indirect cost of the entire system.