7007期中考试答案汇总

## 选择题

1. The following statements are true, except: 100% correct
2. Redundancy level is kept high to reduce the indirect cost of the entire system.
3. Data importance principle is used to decide which data shall be stored and which data shall be discarded. (T)
4. In big data architecture, professional tools are needed to keep the sensitive information safe. (T)
5. Data representation helps to build more meaningful structure for computer analysis. (T)
6. Below are the disadvantages to use traditional file access compared to modern electronic database systems, except: 100% correct
7. Complexity
8. Data manipulation
9. Data order
10. Accessibility
11. The following is false about Hadoop Distributed File System (HDFS), except: 100% Correct
12. Multiple replicas are stored in the same DataNode.
13. NameNode will send heartbeat to DataNode every few seconds to make sure the DataNode is still available. (F)
14. Client reads data from NameNode after DataNode send the necessary information to NameNode.
15. There is only one NameNode in HDFS.
16. Select the statement that are false: 80% Correct
17. Hadoop distributed file system (HDFS) is designed to run across commodity hardware. (T)
18. NameNode actively monitors the number of replicas of a block (T stop here) to make sure the number of replicas is adequate.
19. NameNode send heartbeats to DataNodes to make sure the DataNodes are functioning. (False)
20. Client reads data from NameNode after DataNode send the necessary information to NameNode. (True same to Q3, if this is False, Q3 can’t be C)
21. iii and iv
22. ii and iii
23. i and ii
24. ii and iv
25. Rearrange the following steps that explains the processing of autocoding: 100% Correct
26. Expanding the combinations based on stem forms of words.
27. Matching the allowable variations against the terms listed in the nomenclature.
28. Rearranging the units of the sentence into grammatically permissible combinations.
29. Parsing text into sentences.
30. Allowing for singularities and pluralities of words.
31. Parsing sentences into grammatical units.
32. i – v – ii – iv – vi – iii
33. iv – vi – iii – I – ii – v
34. i – iv – vi – iii – v – ii
35. iv – vi – iii – I – v – ii
36. Select the Hadoop tools that is used for data access: 100% Correct
37. YARN
38. Hive
39. Kafka
40. Ambari
41. ii and iv
42. iii and iv
43. ii and iii
44. i and ii
45. The following is true about classification except: 100% Correct
46. The members of classes maybe highly similar to one another, but their similarities result from their membership in the same class, and not the other way around.
47. Every instance belongs to more than one class. (Exactly one class)
48. 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.
49. In a hierarchical classification, each subclass may have no more than one parent class.
50. Which of the following is wrong about small data? 100% Correct
51. Small data projects are easily replicable.
52. Multiple formats can be used to represent the data.
53. Contain structured data only
54. Usually designed to achieve a specific goal.
55. Information needs to be de-identified because of: 100% Correct
56. Intellectual property.
57. Confidentiality.
58. Data validation. (re-identified, error detection, error correction)
59. Privacy.
60. i, ii and iv
61. ii, iii and iv
62. i, ii and iii
63. i, iii and iv
64. To extract information from text, it is necessary to design structure for it. This may involve: 70% Correct
65. Translating the text to a preferred language.
66. Extracting and normalizing the conceptual terms contained in the sentences.
67. Extracting and standardizing data values from the text.
68. Assigning data values to specific classes of data belonging to a classification system.
69. ii, iii and iv
70. i, ii, iii and iv
71. i, iii and iv
72. i, ii and iii
73. The following is true when using an overly simple classification, except: 100% Correct
74. The classification serves well for specific tasks.
75. The classification includes (exclude) relationships among data objects.
76. The classification includes a “miscellaneous” class.
77. The classes in the classification lack inferential competence.
78. Which of the following is correct about Velocity as one of the V ’S in big data characteristics? 100% Correct
79. Accurateness of the data.
80. Meaning of data in different environment.
81. Analysis of streaming data.
82. Different forms of data.
83. Which of the following is correct about Veracity as one of the V ’S in big data characteristics? 100% Correct
84. Meaning of data in different environment.
85. Different forms of data.
86. Accurateness of the data.
87. Analysis of streaming data.
88. Which of the following is correct about Variety as one of the V ’S in big data characteristics? 100% Correct
89. Analysis of streaming data.
90. Meaning of data in different environment.
91. Accurateness of the data.
92. Different forms of data.
93. Which of the following is correct about Variability as one of the V ’S in big data characteristics? 100% Correct
94. Accurateness of the data.
95. Meaning of data in different environment.
96. Analysis of streaming data.
97. Different forms of data.

问答题：

1. Explain what de-identification means. Then, discuss a logical scenario where de-identification is necessary.

With the promotion of the application of big data in all walks of life, the objects of many big data service scenarios are people. Therefore, much of the sample data used for learning, training, and analysis is personal information. Therefore, these data may contain some personal privacy information such as a person's name, age, address, ID, hobbies, bank account information and so on. The name, ID, address may help data scientists directly locate that the person's data. So, these data need to be de-identified.

De-identification is the process of stripping information from a data record that might link the record to the public name of the record’s subject. It can be understood as a data processing method in which identifiers are processed so that the processed indetifiers cannot directly or indirectly locate to the specific personal information subjects. But it is not achieved by removing an identifier from a data object. For data scientists, de-identification can not only prevent data leakage and protect personal privacy, but also help scientists eliminate interfering information that affects the fairness of data analysis.

For example, Amazon wants to know its users' preferences and shopping habits. the website needs to retrieve customers' data from its database and do data analysis on them. The personal data information include personal identifiers which was collected directly when customers created their accounts. Amazon needs to pre-handle the data through de-identification techniques before analyzing data records to avoid violating their customers' privacy.

1. Explain what re-identification means. Then, discuss a logical scenario where re-identification is necessary.

Re-identification is the reverse process of de-identification. It is the process of matching anonymous information data with publicly available data. Since the data has been de-identified, it may be necessary to re-identify the data when there is sample confusion or data errors when analyzing the data. The re-identification process requires rigorous approval and oversight.

For example: in the analysis of patient data, when the analysis results show that the person has a treatable cancer. However, since the data has been de-identified, the patient himself cannot be directly located. At this time, re-identification is very necessary. Data scientists can entrust data and de-identified records to third parties to ensure data confidentiality.

1. Given “The weather is hot. Air-conditioner need to be turned on.”. Explain how term extraction works to extract indexes from the given sentences, and anticipate the indexes created.

1. Read the first word of the sentence. If it is a common word, delete it. If it is an uncommon word, save it.

2. Read the next word. If it is a common word, delete it and place the saved word (from the prior step, if the prior step saved a word) into our list of terms found in the text. If it is an uncommon word, append it to the word we saved in step one and save the two-word term. If it is a sentence delimiter, place any saved term into our list of terms and stop the program.

3. Repeat step two.

There are two sentences, they contain the words "The", "weather", "is", "hot", "." and "The", "air conditioner", "needs", "to", "be", "turned on", ".".

In the first sentence, "The" is a common word, so delete it. "weather" is kept, "is" is deleted, "hot" is kept, the next sentence separator is ".", so the system stops. So the index of the first sentence is "weather", "hot".

The second sentence, "the" is deleted, "air conditioner" is retained, "needs" is deleted, "to" is deleted, "be" is deleted, "turn on" is retained, the next sentence separator is ".", and the system stops. The index of the second sentence is "air conditioner", "turn on"

1. Discuss TWO (2) important advantages of Hadoop compared to legacy database system, such as relational database in financial sector.

Velocity

The first advantage of Hadoop is Velocity. Hadoop can do real-time processing, offline processing, support pipeline processing, and batch processing. Hadoop is a distributed based storage method. It can 'maps' data wherever it is located on a cluster. The data processing tools can be located on the same server of data so that it can improve the data processing speed. The speed of disk drive addressing time for relational database is much slower than the rate of increase in the transfer rate. If data access is limited by the disk's addressing, it will inevitably cause it to take longer to read or write most of the data. For example, with a financial unstructured data, Hadoop can process terabytes of data in minutes and petabytes of data in hours. Since relational database usually measured in days.

Variability

Another advantage of Hadoop is Variability. Hadoop is a distributed file processing system, so it is a highly scalable storage platform. With Hadoop, large datasets can be stored and provisioned across hundreds of parallel servers. Financial enterprises can run systems on thousands of nodes involving large amounts of data. Since RDB cannot expand to process data, complex databases need to be designed according to different business scenarios.

By the way, the traditional B-tree works well when updating a small portion of the data, but it is not as efficient as MapReduce when updating most of the data because it needs to use sorting/merging to rebuild the database. MapReduce is well suited to handle problems that require analyzing entire datasets in a batch fashion, especially autonomous or real-time analysing. MapReduce is suitable for applications where data is written once and read many times, while relational database is more suitable for datasets that are continuously updated.

1. NameNode assigns the data replica location in such a way to increase the possibility that at least one replica will survive.

Suggest TWO (2) methods that can be used by NameNode to increase data replica survivability. Justify your answer.

Without Namenode, HDFS cannot work. In fact, if the machine running the namenode goes down, the files in the system will be completely lost because there is no other way to reconstruct the files from the blocks located on different datanodes. Therefore, the fault tolerance mechanism of the namenode is very important, and Hadoop provides two mechanisms.

Hadoop can be configured to allow a Namenode to write its persistent state files to a different filesystem. This write operation is synchronous and atomic. A common configuration is to write the persistent state to a remote mounted network filesystem at the same time as writing to the local hard drive.

The second way is to run a secondary Namenode (Secondary Namenode). In fact, the Secondary Namenode cannot be used as a Namenode. Its main purpose is to periodically merge the Namespace image with the edit log to prevent the edit log from becoming too large. Typically, the Secondary Namenode runs on a separate physical machine, as the merge operation takes up a lot of CPU time and memory comparable to the Namenode. The secondary Namenode keeps a backup of the merged Namespace image, which can be used in case the Namenode goes down one day.

1. Explain the importance of distributed computing in Hadoop architecture.

Distributed system refers to dividing a large problem into many small problems and processing them simultaneously in many different computers in the same system. First, users can increase the number of machine nodes at any time according to the task volume and demand to achieve expansion. Administrators can reasonably allocate system resources according to the needs of system tasks. When the demand is high, all machines can be processed at the same time; when the demand is small, some machines can be taken offline to save energy and reduce system wear. Also, because multiple machines have the same services and functions, work does not stop due to unexpected machine downtime. This increases the fault tolerance of Hadoop and reduces the risk of computer failures. Due to its ability to process tasks in parallel, it improves the data processing performance of Hadoop. Significantly improve the data processing speed and cost performance of Hadoop.

1. Differentiate the concept of ontology and classification using at least 5 object classes of your choice. Justify your answer.

classification is the simplest ontologies. Ontologies are constructions that permit an object to be a direct subclass of more than one class.

For example, it is recognized that the Apple mobile phone model is "iPhone se3" by a given image, and "iPhone se3" only belongs to the parent class of "Apple mobile phone model", so it can be understood as classification. Similarly, "iPhone se3" may be A subcategory of "iPhone Model", which can also be a subcategory of "Mobile Phone", "Electronics", "Smart computer", and "2022 New Electronic Products". This can be called ontologies. In this example "iPhone se3" has 5 object parent classes and it can be called a direct subclass of these parent classes. So called ontologies constructions.