

0/15 Questions Answered

Quiz 2

Student Name

Search students by name or email...



Q1 Logistics

0 Points

Please note that unless you have special accommodations for extra time you must submit your answers by 3:55 PM! The system will not kick you out but your submission will be marked as late.

Save Answer

Q2 Z-order

12 Points

Suppose you have 4 2D data items: (2,1), (3,2), (4,2), and (1,3). Put them in Z-order. Order items in ascending order, starting with the least significant bit of the first dimension.

Save Answer

Q3 Amdahl's Law

6 Points

Suppose you have a data science pipeline with 5 operators, which takes 100s to run end-to-end. What is the minimum end-to-end runtime you could ever achieve by further optimizing the operator that currently runs the fastest?

Save Answer

Q4 Python Performance

14 Points

Consider the following python code:

```
#ar1 and ar2 are arrays
#f is an expensive function
out = []
for i in range(0, len(ar1)):
    for j in range(0, len(ar2)):
        if f(ar1[i]) == f(ar2[j]):
            out.append((ar1[i], ar2[j]))
```

List 2 ways you could optimize this code to make it run faster. Consider algorithmic changes, parallelization, etc. Assume the order of the output does not matter.

Save Answer

Q5 Partitioning and Joins

18 Points

Suppose you frequently join three tables, S, T, and R with join predicates $S.a = R.b$ and $T.c = R.d$. These tables are filtered differently in different queries, but for most queries the filtered versions of S are about 1 MB, filtered versions of T are about 100 MB, and filtered versions of R are about 10 GB. Each record in each database is 100 bytes. Assume R is a fact table and S and T are dimension tables (i.e., this is a snowflake schema) and you are doing primary key-foreign key joins.

You are running on a 4 node cluster, with each node having 8

GB of RAM and 1 processor. Your network can transfer 10 MB / sec and nodes can compare 10 M records/sec.

Which of the following pre-partitioning and join strategies would you recommend, and why?

Hash partition S on S.a, T on T.c, and R on R.d. First locally join partitions of T with R via a hash join, then shuffle the result on R.b across the four nodes and join with partitions of S on S.a.

Hash partition S on S.a, T on T.c and R on R.b. First locally join partitions of S with R via a hash join, then shuffle the results on R.d across the four nodes and join with partitions of T on T.c.

Hash partition R on R.d and T on T.c. At query time, replicate the filtered version of S across all four nodes. First locally join partitions of R with T via a hash join, then join each partition locally with S.

Hash partition R on R.b and S on S.a. At query time, replicate the filtered version of T across all four nodes. First locally join partitions of R with S via a hash join, then join each partition locally with T.

Explanation:

Save Answer

Q6 Parquet

12 Points

How large would the Parquet encoding of the following array of 8-byte integers be, in bytes?

20 21 20 19 18 20 21 21 18

Assume that Parquet uses the first element as the reference value, which is stored in parquet using 8 bytes. Subsequent input values are processed in 64-byte blocks. Each compressed block is accompanied by an 8-byte min delta and a 1-byte (max)bits value. Explain your answer.

72 bytes

41 bytes

20 bytes

23 bytes

Save Answer

Q7 Lying with Statistics/Visualizations

12 Points

Name at least 3 problems with the visualization below and explain briefly how the visualization is misleading:



Save Answer

Q8 Multiple Hypothesis Testing

14 Points

You just became the Head Data Scientist at a new online gaming startup. Your first job is to determine which of your 1000 early users are most likely to buy virtual coins. For each user you know the following attributes:

1. Age
2. Bachelor's degree (yes/no)
3. Master's degree (yes/no)
4. PhD degree (yes/no)
5. Facebook user (yes/no)
6. Instagram user (yes/no)
7. TikTok user (yes/no)
8. Spent over 1h in game (yes/no)
9. iPhone user (yes/no)
10. Purchase amount

Q8.1 P-value I

3 Points

You have a hunch that age impacts how much your users spend. As a first step, you divide the dataset into young (under 21) and old users (above or equal to 21) and you perform a two-sided t-test with a significance value of $\alpha = 0.05$.

If you know that you end up rejecting the null hypothesis after this test, what does the result mean?

The chances are greater than 1 in 20 that a difference would be found if the study were repeated.

The probability is less than 5% that a purchase difference this large could occur by chance alone.

The probability is less than 5% that a purchase difference this large or larger could occur by chance alone.

With >95% probability older people have an average purchase power of \$13 whereas younger people of \$15.

Save Answer

Q8.2 P-value II

1 Point

Assume you perform the same two-sided t-test as in the previous question.

If you instead know that the t-test gives you a value of $p = 0.021$, do you reject or accept the null hypothesis?

Reject the null hypothesis

Accept the null hypothesis

Save Answer

Q8.3 Bonferroni Correction

5 Points

Encouraged by the result, you test 4 additional attributes to see how they impact sales. Specifically, you also use a two-sided t-test for "Facebook user", "Instagram user", "TikTok user", and "Spent over 1h in game".

The results for all the tests are below (including the one for age):

P-value	Feature Tested
0.005	Spent over 1h in game
0.006	Instagram user
0.015	TikTok User
0.021	Age (young/old)
0.051	Facebook user

However, performing this many tests increases the risk of having at least one type I error. Hence, you decide to do a Bonferroni correction to keep the family-wise error rate at 5%. With the Bonferroni correction, which test differences would be considered significant?

☐ Spent over 1h in game

☐ Instagram user

☐ TikTok User

☐ Age (young/old)

☐ Facebook user

Save Answer

Q8.4 Correlation Matrix

5 Points

You decide to do one more thing and create a correlation matrix of all 10 features to quickly look at all potential correlations in the dataset as shown below:

Feature Number	Feature Number									
		2	3	4	5	6	7	8	9	10
	1	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$	$C_{1,6}$	$C_{1,7}$	$C_{1,8}$	$C_{1,9}$	$C_{1,10}$
	2		$C_{2,3}$	$C_{2,4}$	$C_{2,5}$	$C_{2,6}$	$C_{2,7}$	$C_{2,8}$	$C_{2,9}$	$C_{2,10}$
	3			$C_{3,4}$	$C_{3,5}$	$C_{3,6}$	$C_{3,7}$	$C_{3,8}$	$C_{3,9}$	$C_{3,10}$
	4				$C_{4,5}$	$C_{4,6}$	$C_{4,7}$	$C_{4,8}$	$C_{4,9}$	$C_{4,10}$
	5					$C_{5,6}$	$C_{5,7}$	$C_{5,8}$	$C_{5,9}$	$C_{5,10}$
	6						$C_{6,7}$	$C_{6,8}$	$C_{6,9}$	$C_{6,10}$
	7							$C_{7,8}$	$C_{7,9}$	$C_{7,10}$
	8								$C_{8,9}$	$C_{8,10}$
	9									$C_{9,10}$

Based on the correlation matrix you also discover that a Bachelor degree seems to impact how much time a user spends in the game. You do a one-sided t-test afterwards and the p-value is 0.003. Controlling the family-wise error rate at 5% using Bonferroni, would you consider this as significant enough to reject the null-hypothesis? How would it impact your previous results? Explain your answer.

Save Answer

Q9 Compression Techniques

12 Points

Which compression technique would you choose for each of the following types of data? If you make any assumptions, explain them in the box below each part.

Q9.1

3 Points

A column in a database of MIT faculty, sorted by name, indicating the year in which they won a Turing Award.

Delta encoding

Run-length encoding

Only store non-null tuples

Dictionary encoding

Save Answer

Q9.2

3 Points

Quality control sensor data owned by a soda company, stored in chronological order, indicating the exact fill level of each bottle as it leaves the production line.

Delta encoding

Run-length encoding

Only store non-null tuples

Dictionary encoding

Save Answer

Q9.3

3 Points

A column in MIT Medical's student database, sorted by MIT student ID, indicating the manufacturer of the COVID-19 vaccine received by each student.

Delta encoding

Run-length encoding

Only store non-null tuples

Dictionary encoding

Save Answer