SAN JOSÉ STATE
UNIVERSITY

CMPE272 Mid-Term Exam March 13, 2017 6:15-8:00PM SET-A

Student ID: **Student Name:**

Directions (Please read carefully): Please make sure you write your Student ID and Name. You are required to answer all questions. There is only one correct answer for multiple choice questions, unless you are asked to choose specific number of right answers. Circle your answer for easy readability.

Multiple Choices (Each carry 2 marks):

1. Suppose that you have an application whose behavior depends on the environment variable LOG_DIR. Which of the command line may be used in a bash shell to configure the application?
   a. Export $LOG_DIR=/tmp; echo $LOG_DIR;
   b. Set LOG_DIR=/tmp
   [✓ c. LOG_DIR=/tmp; export LOG_DIR]
   d. None of the above
2. The SQL WHERE clause
   a. limits the column data that are returned.
   [✓ b. limits the row data are returned.]
   c. Both A and B are correct.
   d. Neither A nor B is correct.
3. Which of the following statements are **true**?
   [✓ a. Linux uses a technique called demand paging.]
   b. Linux uses a Most Frequently Used(MFU) technique to swap a page.
   c. Both (a) and (b)
   d. None of the above
4. Which of the following statements are **true**?
   a. Interrupt is a signal sent to the hardware from the operating system.
   [✓ b. SIGINT is generated when you hit 'Ctrl+C']
   c. Both (a) and (b)
   d. None of the above
5. *Select name from instructor where dept name = 'Chemistry' order by name;* By default, the order by clause lists items in ______ order.
   a. Descending
   b. Any
   [✓ c. Ascending]
   d. Same
6. In ____________ databases, each object stored is mapped with "nodes" and "edges".
   a. Time series
   b. GIS
   [✓ c. Graph]
   d. Columnar

7. ______________ is the process of steps that will identify, for elimination, redundancies in a database design.
    a. Serialization
    b. De-normalization
    c. Normalization
    d. Decluttering
8. ______________ is the process of trying to improve the read performance of a database, at the expense of losing some write performance, by adding redundant copies of data or by grouping data.
    a) Serialization
    b) De-normalization
    c) Normalization
    d) Decluttering
9. The namenode knows that the datanode is alive by using a mechanism known as
    a. Datapulse
    b. H-signals
    c. Active-pulse
    d. Heartbeats
10. Which of the following is **not** a goal of HDFS?
    a. Fault detection and recovery
    b. Handle huge dataset
    c. Prevent deletion of data
    d. Provide high network bandwidth for data movement

True/False Questions (each carry 2 marks):

1. The following query is syntactically correct **True / False**

SELECT WORKDEPT, AVG(SALARY) FROM EMPLOYEE WHERE AVG(SALARY) > 20000 GROUP BY WORKDEPT HAVING COUNT(*) 3 ORDER BY 2 DESC

2. Signals are types of Inter process communication (IPC) mechanisms which are bi-directional. **True / False**
3. Linux “named” pipes operate on LIFO (Last In First Out) basis. **True / False**

4. val textFile = sc.textFile("/user/emp.txt") This statement creates an RDD but the file itself is not loaded at this point. **True / False**
5. When volume of data increases, cardinality also increases. **True / False**

6. RDDs in Spark are immutable. **True / False**

7. You can have only mapper function in your map reduce job by turning the reducer function off. **True / False**
8. HDFS replication factor is used to make a copy of the data (i.e) if your replicator factor is 2 then all the data which you upload to HDFS will have a copy.**True / False**

9. Default HDFS block size is fixed and it can never be changed. **True / False**

10. The Hadoop (HDFS) balancer moves blocks around from one node to another to try to make it so each datanode has the same amount of data **True** / **False**
11. Infrastructure as a Service (IaaS) describes a distribution model in which applications are hosted by a service provider and made available to users. **True** / **False**
12. A foreign key is designed into a table to define relationships between columns **True** / **False**

13. RDD will use cache() functions to cache the operation only if data fits into memory **True** / **False**
14. Column store database offer optimal write time and abundant reading overhead for retrieval of subset queries. **True** / **False**
15. RDDs track *lineage* information that can be used to efficiently re compute lost data **True** / **False**

Short Answer Questions (2 Marks each)

1. What is the difference between UNION and UNION ALL?

UNION will omit duplicate records whereas UNION ALL will include duplicate records.

2. Correct the query below so it executes properly. *SELECT Id, YEAR(BillDate) AS BillYear FROM Invoices WHERE BillYear >= 2001;*

SELECT Id, YEAR(BillDate) AS BillYear FROM Invoices WHERE YEAR(BillDate) >= 2001;

3. How are the “on-delete-cascade” and “on-delete-restrict” referential integrity delete-rules for foreign keys, different from one another?

On-delete-cascade will delete all referencing rows for foreign keys
On the other hand, in on-delete-restrict, if a row with values were found the delete will fail.

4. What does 2>&1 mean?

Redirect STDERR to STDOUT. Used for logging both error and standard output.

5. What does Seconds(1) mean in the Spark streaming code example below:
val ssc = new StreamingContext(args(0), "NetworkWordCount", Seconds(1)

Time interval at which streaming data will be divided into batches

6. Name 2 Database SCALAR functions

Pick any from the list: DECIMAL
SUBSTR
COALESCE
VALUE
YEAR
MONTH

DAY
DAYS

7. From which SQL statements may primary and foreign keys be defined?

CREATE TABLE and ALTER TABLE

8. Write a mapper/reducer function for counting words in a file.

9. Write the Linux command to display the 'file' contents per page.

Cat <file name> | more

10. Write a SQL query for finding the second highest salary in an employee table containing "salary" and "id" fields.

SELECT Salary FROM (SELECT Salary FROM Employee ORDER BY salary DESC LIMIT 2) AS Emp ORDER BY salary LIMIT 1

11. Write a SQL query to list Last names of employees whose age is either 25, 35 or 45 (Hint: use the IN predicate)

SELECT LNAME FROM EMPLOYEE WHERE AGE IN (25,35,45) ORDERBY AGE, LNAME

12. What is the first process to execute when linux kernel is loaded? What could happen if kernel cannot execute this process?

Init()
Goes into panic mode

13. Explain why query optimization is necessary.

Much more efficient, enhance query performance by producing only necessary tuples to reduce I/O reading time

14. How does MapReduce improve the reliability of distributed systems?

Map reduce improve the reliability by replication in which each datanode is replicated. A datanode has 3 replicas one on clients machine one on a different rack and the 3rd on the same rack as 2nd but on different node

15. Mention 2 advantages of HDFS's data block approach.

Simplifies replication, providing fault tolerance and reliability, and shields users from storage subsystem details.

16. How does name-node help in recovery from failed data-nodes in the HDFS?

Data blocks are replicated across multiple nodes so when one node fails it will just use another node.

17. How does adding more data-nodes cause clusters to be 'unbalanced'?

When node is added HDFC does not transfer the data automatically . It writes data on new node as it comes. So the old node has more data compared to the new node causing unbalancing. Removed by using Balancer .

18. What is sparkContext in Apache spark? Can I create RDD without creating sparkContext?

Mark the entry point for spark functionality. **Rdd cannot be created without sparkContext**

19. What is the significance of 'copy-on-write' in the context of fork() system call?
20. What is the main difference between mutex and semaphore.

In the case of mutex, only the thread that locked or acquired the mutex can unlock it. In the case of a semaphore, a thread waiting on a semaphore can be signaled by a different thread.

21. Mention any 2 forms of IPC in System V

Semaphore, Shared Memory, Message Queues

22. Give few examples of "Transformations" in RDD operations.

Transformation is creating new datasets from the existing datasets. example map, distinct, groupkeys , join, union.

23. When do you use Hive versus Pig?

It is best used for batch jobs over large sets of immutable data

24. Mention 2 use-cases for GIS databases.

Finding Use Cases: Identify areas with high crime rates to determine where to assign more resources.

Finding Locations: Find best location for new warehouse by identifying lots that are vacant, at least five acres.

25. Why is the startup and shutdown time of containers faster than VMs?

Containers do not need multiple instances for different process they use the same kernel for for all the process that's why they are fast.