Week 5 Quiz Solution

1. Suppose we have four tables Employee, Department, Location, Skills, table. We fire a query in Hive:
   SELECT * FROM Department JOIN Employee
   ON (Department.emp_id =Employee.emp_id) JOIN Location
   ON (Employee.location_id =Location.location_id)
   JOIN (Employee.skill_code = Skills.skill_code)
   How many MapReduce Jobs will run in this case?
   a. ***3**
   b. 4
   c. 2
   d. 1

   Explanation: Total 3 columns are involved in join condition(emp_id , location_id, skill_code) hence there will be 3 MT jobs

2. If the above query is modified as
   SELECT * FROM Department JOIN Employee
   ON (Department.emp_id =Employee.emp_id) JOIN Location
   ON (Employee.location_id =Location.location_id)
   JOIN (Employee.emp_id = Skills.emp_id).
   How many Mapreduce Jobs will run in this case?
   a. 3
   b. 4
   c. *2
   d. 1

   Explanation: Total 2 columns are involved in join condition (emp_id , location_id) hence there will be 2 MR jobs

3. Which of the following is incorrect (you can select multiple answers)
   a. partitioning->bucketing
   b. partitioning->partitioning->bucketing
   c. *partitioning->bucketing->bucketing
   d. *partitioning->bucketing->partitioning->bucketing
   e. partitioning->partitioning

   Explanation: C & D
   Bucket is a file. File cannot contain file or directory.

WEEK 5 QUIZ SOLUTION

4. Suppose you have a table employee with these columns: E_id E_name E_department E_state
   Suppose you run the query on:
   E_id 1000000 times a day
   E_state 10000 times a day
   E_department 1000 times a day
   Which column would you partition on? (multiple can be correct)"
   a. ***E_state**
   b. E_id
   c. ***E_department**
   d. does not matter

   Explanation: A & C is perfect match of partition. On E_id, there are more queries executed in a day but creating partition wont help because partition is a directory and it should hold multiple records. 1 partition = 1 row is not an ideal solution.

5. Which one of the partitioning techniques is faster?
   a. ***static**
   b. Dynamic
   c. depends on the data
   d. both are of the same speed

   Explanation: static is faster because it does not check for partition values. We load every partition data separately yourselves.

6. Advantage of map side join
   a. no shuffling involved
   b. no reducer required
   c. no sorting required
   d. ***All of the above**

   Explanation: In Map-Side join only mapper is executed hence shuffle, sort and reducer does not play the role.

7. The advantages of creating partitions in a table are
   a. ***It makes querying faster**
   b. Effective storage and memory utilization
   c. compression techniques
   d. performant engine

   Explanation: By creating partition, we get faster query result because it does not check for all data sequentially. It will jump to particular partition and then search sequentially.

WEEK 5 QUIZ SOLUTION

8. Let's say you have the sales data of an e-commerce firm, for the year 2016, stored in a Hive table. The columns in the table are as follows: Order_ID Date (dd/mm/2016) Month Customer_ID Customer_city Product_ID Sales_in_INR There are a total of 2,000,000 unique orders, 365 days, 90,000 unique customers, 25 cities, and 15,000 unique products.
   You frequently need to do the following:
   Compare the average sales in Chennai, Hyderabad, and Mysore
   Compare the average sales across each month
   Fetch the order history of a particular customer
   Compare the total sales in INR of all the products
   You notice that running queries on the entire dataset take a lot of time, and you decide to use partitions and buckets.
   Now, answer the following questions.
   Which of the following columns should you NOT use as a partition key? More than one options may be correct."
   a. Customer_city
   b. Month
   c. *order_id
   d. *customer_id

   Explanation: C & D. Pay attention it is asking for NOT. Customer_id and order_id is not frequently queried.

9. Let's say that you created partitions using 'Customer_city' as the key. You notice that even after partitioning, queries such as 'fetch the order history of customers in a particular city', which you need to run very often, are too slow.
   Which one of the options given below should you take to make such queries faster?"
   a. create another level of partitioning using "Customer_ID" as key
   b. Use a different partition key
   c. *Create about 100 buckets using the column 'Customer_ID'
   d. Create another level of partitioning using "Month" as key

   Explanation: There is a partition on customer_city column, still some cities take long time to fetch customer data. Customer_id is unique, high cardinality hence creating bucket is an ideal solution.

WEEK 5 QUIZ SOLUTION

10. Suppose that you created 100 buckets using the column 'Customer_ID', this will result in a hash function being specified, which creates 100 buckets. Now, as the database grows with time, which one of the results below will you see:
    a. The number of buckets will increase automatically
    b. *The number of data entries in individual buckets will increase automatically
    c. First the number of data entries in individual buckets will increase but post a limit the number of buckets will also start increasing
    d. we can't say

    Explanation: When we create buckets, the number of buckets always remain same. Data keeps on coming and gets appended to respective bucket. Bucket is a HDFS file, if it grows then it will keep on creating different data blocks.

11. Which of these clauses when used in a query, ensure non-overlapping values among reducers? Multiple options can be correct.
    a. Sort by
    b. *Distribute by
    c. *Distribute by Sort by
    d. *Cluster by

12. If number of reducers is set to 1, using hive property, sort by clause when used in a select query, on a column will give same output as using order by clause in the query?
    a. *TRUE
    b. FALSE

    Explanation: Sort by does local sorting (for every reducer). Order by does global sorting because it always uses one reducer. After setting reducer to 1 explicitly then sort by will also use one reducer and hence output of order by and sort by will be same.

13. Ties are assigned same rank, and ranks are consecutive in case of?
    a. *dense_rank()
    b. rank()
    c. row_number()
    d. All of the above

WEEK 5 QUIZ SOLUTION

14. Consider we are creating bucket on customer_id column in customer table. we created 8 buckets. In which bucket customer_id 25 will go (consider bucket id from 0 to 7)
    a. ***bucket id 1**
    b. bucket id 0
    c. bucket id 5
    d. bucket id 4

    Explanation: 25%8=1. Customer_id%Buckets=Bucket-Number.
    Another e.g : 24%8=0 Record with ID 24 will go to bucket 0

15. Suppose your most frequent query is:
    select * from customers where state=? and customer_id =?
    Then what's the best way to structure your table
    a. partitioning on both state and customer_id column
    b. bucketing on both state and customer_id column
    c. ***partitioning on state and bucketing on customer_id**
    d. bucketing on state and partitioning on customer_id

    Explanation: State has low cardinality hence partition. Customer_id has high cardinality hence we can use hash function and divide the records into buckets.

16. TrendyTech uses two large tables in hive named weeks (4 Buckets) and weeksClicks(8 Buckets). Both tables have weekID in common but are not sorted. Which type of join is best suited in this scenario.
    a. Map Side Join
    b. ***Bucket Map Join**
    c. Sort Merge Bucket Join

    Explanation: Bucket Map join rules are as follows
    - Both tables can be large.
    - Both tables have common join column
    - Both tables have buckets and in proportion
    - Data inside both tables are not sorted

17. John is using windowing function with partition by and does not give order by clause then what will be the default setting of window
    a. ***rows between unbounded preceding and unbounded following**
    b. rows between unbounded preceding and current row
    c. rows between current row and unbounded following
    d. rows between 1 preceding and 1 following

    Explanation: If order by clause is not specified then A else B

WEEK 5 QUIZ SOLUTION

18. Which of the following is true regarding partition in Hive?
    a. ***Partition can be done where cardinality is less**
    b. Partition can be done where cardinality is more
    c. ***Partition creates directories**
    d. Partition creates files

Explanation: A & C.
Why B is wrong?
If we create partition on column with high cardinality. e.g. Customer_Id then it will create too many directories and every directory will just have one record, hence no performance benefit.

19. Which of the following is not an advantage of map side join?
    a. Improves processing time
    b. ***Increases data transfer between nodes in the cluster**
    c. Reduce shuffle and sort time

Explanation: B, it reduces the data transfer between nodes in the cluster

20. Which ranking function assigns unique number regardless of the value
    a. dense_rank()
    b. rank()
    c. ***row_number()**
    d. All of the above