

Modern Big Data Analysis with SQL

Coursera Specialisation (Offered by Cloudera)

Course-3: Managing Big Data in Cluster and Cloud Storage

Week-5: Honours

Question-1: What is a view in Hive or Impala? Check all that apply.

Answer-1: A way to simplify access to a table for operations that are frequently repeated
A way to restrict access to certain columns

Question-2: What happens in the file system when you drop a (non-materialized) view with Hive or Impala?

Answer-2: No data is deleted

Question-3: Which of the following can help you find out how Hive or Impala will execute a query? Check all that apply.

Answer-3: Add EXPLAIN at beginning of the query, execute it as you would a normal query
- Enter the query in Hue and click the folded map icon

Question-4: Following is a query and the first portion of the Hive execution plan for it. Which stages will be done before Stage-2? Check all that apply.

```
SELECT country, empl_id, COUNT(c.cust_id) AS num
```

```
FROM customers c
```

```
JOIN orders o ON (c.cust_id = o.cust_id) GROUP BY country, empl_id;
```

Answer-4: Stage-5

Question-5: Under what circumstances might it be a good idea to use partitioning?

Check all that apply.

Answer-5: The data is very large and queries take too long to complete

- Most queries filter the data based on one or two specific columns
- The data files are already divided into subdirectories whose names specify a value not given in the files

Question-6: Which column would make a good partition column?

Answer-6: A categorical column with a reasonable number of different values

Question-7: Which complex data type represents key-value pairs, with all keys having the same data type and all values having the same type? (However, a key and its associated value may have different types.)

Answer-7: MAP

Question-8: True or false: The fields in a STRUCT may have different data types.

Answer-8: True

Question-9: When using Impala to query tables with complex columns, which file formats can you use to store the data? Check all that apply.

Answer-9: Parquet

Question-10: Which of the following statements accurately describe storage engines like HBase and Kudu? Check all that apply.

Answer-10: They encapsulate data storage, exposing a high-level interface to the data

- They abstract away the details of how the data is stored using specific file formats