Quiz 5: DSC 208 Data Management for Analytics

Questions and Explanations

Question 1. Which of the following systems capabilities is a key differentiator of data "lakehouses" as against data warehouses?

a) Integration with business intelligence tools

b) SQL as a user-facing language

c) Loose coupling of file format with query stack

d) Automated query optimization

Answer: c) Loose coupling of file format with query stack

- *Explanation:*
  - 'Data warehouses' traditionally tightly couple their storage format (often proprietary or highly optimized for their specific query engine) with the query engine itself.
  - 'Lakehouses', however, combine the flexibility of data lakes (which store data in open formats like Parquet, ORC in object storage) with the analytical capabilities of data warehouses. This means they offer a 'loose coupling of the file format with the query stack'. You can use various query engines (Spark SQL, Presto, Hive, etc.) on the same underlying data files in the data lake, without being locked into a single vendor or format.
  - Options a), b), and d) are common features of both modern data warehouses and lakehouses, so they are not key differentiators *between* them.

Question 2. What was the main novel technical capability of Spark relative to parallel RDBMSs when it was introduced?

a) Optimized query execution

b) Scales to multi-node clusters

c) High-level querying/API

d) Lineage-based fault tolerance

Answer: d) Lineage-based fault tolerance

- *Explanation:* When Spark was introduced, while it offered capabilities like optimized query execution and high-level APIs similar to what parallel RDBMSs aimed for, and scalability to multi-node clusters (which MapReduce also did), its truly novel technical capability was ‘lineage-based fault tolerance‘ through Resilient Distributed Datasets (RDDs). Spark could reconstruct lost partitions of data by replaying the transformations (the ”lineage”) that produced them, rather than relying on full data replication or checkpoints, leading to significant performance gains, especially for iterative algorithms and interactive queries, compared to the disk-heavy MapReduce.

Question 3. Which of the following feature engineering steps requires only a Map-only job to fully scale using MapReduce?

a) Whitening

b) Pairwise feature interactions

c) One-hot encoding

d) All of the three

Answer: b) Pairwise feature interactions

- *Explanation:*
  - A ‘Map-only job‘ in MapReduce means that each input record can be transformed independently without needing to aggregate or combine information from other records.
  - ‘Pairwise feature interactions‘ (e.g., creating a new feature by multiplying two existing features $F1 \times F2$) can be done entirely within the Map phase. For each input record, you simply access the values of $F1$ and $F2$ for that record and compute the new feature. This is a local transformation per record.
  - ‘One-hot encoding‘ often requires knowing the *global* vocabulary or distinct values for a categorical feature to create the appropriate number of binary columns. While mapping can start, a Reduce step (or a global pass beforehand) is typically needed to collect all unique categories.
  - ‘Whitening‘ (or Z-score normalization) requires computing global statistics like the mean and standard deviation of a feature across the *entire dataset*. This necessitates a Reduce step to aggregate these statistics before the actual transformation can occur.

  Therefore, only ‘pairwise feature interactions‘ can be fully scaled using only a Map-only job.

Question 4. Which of the following is not a major type of data cleaning tasks?

a) Local edits to cell values

b) Reconciling across tuples

c) Synthesizing table values

d) Reconciling values in a column

Answer: c) Synthesizing table values

- *Explanation:*
  - ‘Data cleaning‘ primarily involves identifying and correcting errors, inconsistencies, and inaccuracies in data.
  - ‘Local edits to cell values‘ (e.g., correcting typos, standardizing formats within a single cell) are core cleaning tasks.
  - ‘Reconciling across tuples‘ (e.g., entity resolution, deduplication where information from multiple records is combined or corrected) is a major cleaning task.
  - ‘Reconciling values in a column‘ (e.g., ensuring consistency of units, handling missing values in a column) is also a key cleaning activity.
  - ‘Synthesizing table values‘ refers to generating new data or inferring values that were not originally present, often for purposes like data augmentation or imputation, but it’s not typically classified as a ”cleaning” task, which focuses on rectifying *existing* erroneous data.

Question 5. What is an outlier?

a) A data point that is duplicated in the dataset

b) A data point that is missing from the dataset

c) A data point that deviates from the norm

d) A data point that has a value of zero

Answer: c) A data point that deviates from the norm

- *Explanation:* An ‘outlier‘ is a data point that significantly differs from other observations. It’s an observation that lies an abnormal distance from other values in a random sample from a population. Outliers can indicate variability in a measurement, experimental errors, or a novelty. They are distinct from duplicates (repeated entries), missing values (absent data), or zero values (which can be normal).