DSC 208 Module 2 and 3 Questions

Module 2 Questions

1. **Name three common forms of structured data with application examples.**
    - **Relational Data:** Ubiquitous; e.g., a transactional database.
    - **Data Frame Data:** Common in recommendation systems and tabular analysis.
    - **Matrix and Tensor Data:** Used in statistical and scientific computing.
2. **What is Parquet? Explain one pro and one con of using Parquet versus CSVs.**
    - **Definition:** Parquet is a columnar, compressed file format for structured and semi-structured data.
    - **Pro:** Smaller file size and faster access via column pruning.
    - **Con:** Not human-readable or easily editable.
3. **What is a data lake? How is it different from an RDBMS?**
    - **Data Lake:** A file system storing diverse native-format files; supports direct file access.
    - **Difference:** RDBMS uses query stacks and may not allow direct file access; data lake does.
4. **Explain two reasons why data acquisition can be challenging and how to mitigate them.**
    - **Heterogeneity:** Mitigate by assessing source necessity.
    - **Access Control:** Mitigate by learning and adhering to access policies.
    - **Manual Errors:** Address with robust validation and error handling.
5. **Explain two best practices for data reorganization or preparation.**
    - **Automation:** Use workflow tools.
    - **Documentation:** Maintain shared, clear documentation.
6. **Explain one pro and one con of programmatic (over)labeling.**

- **Pro:** Increases productivity and reduces costs.
- **Con:** Requires coding skill; not universally applicable.

7. **Name a data privacy law that affects many web companies.**
   - GDPR and CCPA/CPRA.
8. **What is data governance? Why should we track it?**
   - **Definition:** Management of data availability, usability, integrity, and security.
   - **Purpose:** Ensures auditability, compliance, and continuity.

Module 3 Questions: Semi-Structured Data and Graph Databases

1. **Two ways semi-structured data models differ from relational data:**
   - Schema flexibility
   - Heterogeneous records
   - Nested structures
2. **Two applications for semi-structured data models:**
   - User profile management
   - Data exchange and integration
3. **Difference between XML and JSON:**
   - XML uses tags; JSON uses key-value pairs and is less verbose.
4. **Difference between a tag and an attribute in XML:**
   - **Tags:** Define elements and can contain sub-elements.
   - **Attributes:** Provide metadata and are atomic.
5. **Basic form of an XQuery statement:**

```
FOR $var IN ...
[LET $var := ...]
[WHERE condition]
RETURN expression
```

6. **How XQuery resembles sequence syntax:**
   - WHERE clause acts like a predicate; path expressions used for selection.

7. **One way JSON is better than XML:**
    - Simpler syntax; easier to parse and read.
8. **Motivation for key-value/NoSQL stores:**
    - Needed scalability, availability, and schema flexibility for large-scale web applications.
9. **Two major types of graph processing systems:**
    - OLTP-like (Transactional)
    - OLAP-like (Analytical)
10. **Key benefit of GraphX over custom graph DBMSs:**
    - Integrated with Spark; no separate system required; supports hybrid relational-graph operations.

Quiz 2

- **Primary key (A, B) implies A = B: False**
- **Query Result Tuple: (4,3,1)**
- X^+ **is always a superkey: True**

Quiz 3

- **Hash indexes for 4 attributes: 18**
- **Predicate supported by both hash and B+ tree: Equal to**
- **Benefit of declarativity: All of the three**
- **Intersection tuple in R ∩ S: (1,2,3)**
- **Theta-join result tuple: (1,2,2,4,6)**
- **Selectivity of NOT(Stars ¿ 3.0): 40%**