Research for SQL & Data Modeling Sprint

1. Why is structured data important in data science pipelines?

Structured data is essential in data science pipelines because it provides precision, consistency, and efficiency throughout the analytical process. Its well-defined tabular format (organized into rows and columns) enables fast and accurate data operations such as filtering, joining, querying, and aggregation—tasks that are foundational in data analysis. Tools like SQL, pandas, and business intelligence platforms are specifically designed to work efficiently with structured data, making it easier to manage and analyze large datasets [1].

Furthermore, structured data improves the accuracy and reliability of machine learning models. Clean, consistently formatted data can enhance model performance by up to 70% and improve forecasting accuracy by around 30%. It also integrates smoothly with machine learning frameworks such as IBM SPSS, supporting real-time analytics and decision-making in automated pipelines [2].

References:

[1] Muthukkaruppan, S., et al. (2021). *Data Systems for Machine Learning*. ACM Computing Surveys from https://dl.acm.org/doi/10.1145/3267338

[2] MetricsRule. (2023). *Impact of Structured Data on Predictive Accuracy and Forecasting*. Retrieved from https://metricsrule.com

2. What role does data modeling play in preparing data for analysis or machine learning?

Data modeling is a fundamental process for organizing and structuring data clearly, making it easier to store, access, and analyze. One common method of data modeling is creating an **Entity Relationship Diagram (ERD)**, which visually represents entities (such as students, courses) and their relationships in an organized way.

An ERD acts as a blueprint that helps integrate data from multiple sources and guides how data should be transformed and loaded during ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform) workflows. This makes data pipelines more reliable, scalable, and easier to manage.

In machine learning, data modeling and ERDs help reveal important relationships and groupings between variables, simplifying tasks like normalization and aggregation in feature engineering. A well-designed data model also improves query performance and makes data easier to understand, which supports better analysis and decision-making.

Reference:

Wikipedia. (2023). Data modeling. Retrieved from https://en.wikipedia.org/wiki/Data_modeling

3. How do relational databases support scalable and clean data practices in real-world data science projects?

Relational databases are fundamental in data science for ensuring data integrity, scalability, and efficient management. Here's how they contribute:

1. Structured Data Modeling with Normalization
   Relational databases organize data into structured tables with predefined relationships, applying normalization techniques to minimize redundancy and enhance data integrity. This structured approach facilitates efficient data storage and retrieval, crucial for large-scale data analysis [1].

2. Enforcement of Data Integrity Constraints
   They implement integrity constraints such as primary keys, foreign keys, and domain restrictions to maintain data accuracy and consistency. These constraints prevent anomalies and ensure reliable data relationships, which is essential for accurate data analysis and decision-making [2].

3. Efficient Data Retrieval and Scalability
   Relational databases support efficient data retrieval through indexing and optimized query processing. They can handle increasing data volumes and complex queries, making them scalable solutions for growing data science projects [3].

References

[1] Number Analytics. (2023). Relational Databases 101. Retrieved from https://www.numberanalytics.com/blog/relational-databases-data-science-ultimate-guide

[2] Wikipedia. (2023). Data Integrity. Retrieved from https://en.wikipedia.org/wiki/Data_integrity

[3] Wikipedia. (2023). Relational Database. Retrieved from https://en.wikipedia.org/wiki/Relational_database

4. Why is SQL still considered a foundational skill even with tools like Python and Pandas?

SQL remains essential because it is **faster and more efficient** for working with large datasets directly inside the database. It uses indexes and query optimizers to handle **millions of rows** much faster than Pandas, which loads all data into memory and can slow down or crash with big data sets [1].

Also, SQL keeps data **well-structured at the source**, while Python and Pandas are better for advanced analysis after you extract just what you need [2].

In practice, the best approach is to use SQL to filter and prepare the right data, then switch to Pandas for deeper processing.

References

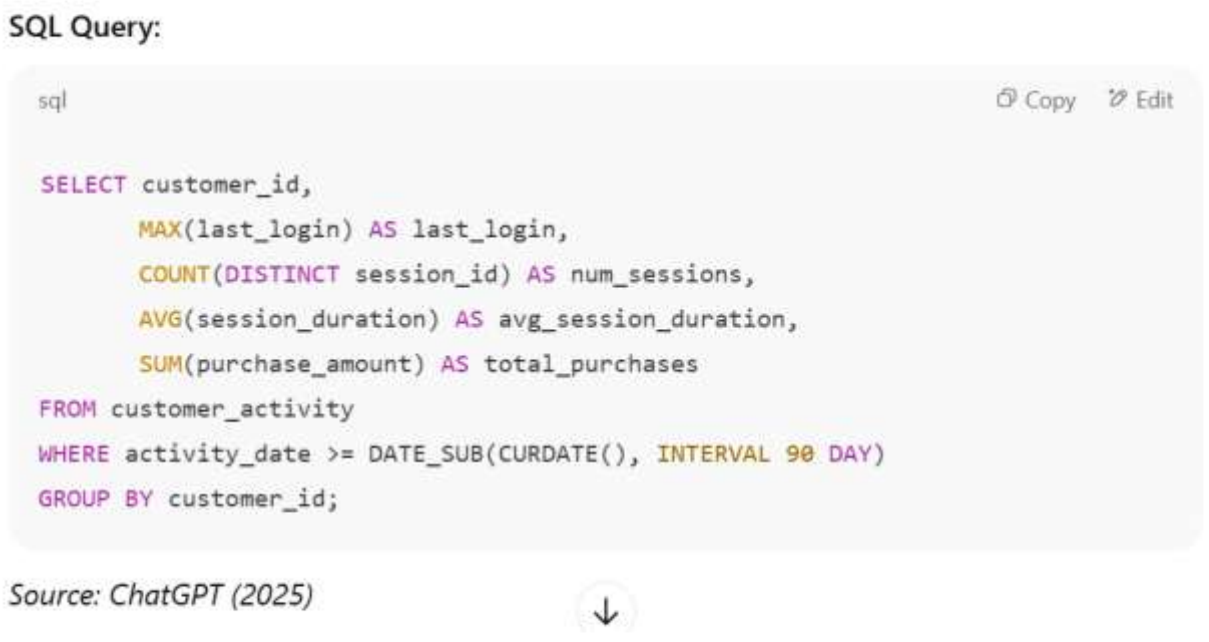
[1] Saturn Cloud. *Pandas vs SQL Speed: A Comparison*. Available at: https://saturncloud.io/blog/pandas-vs-sql-speed-a-comparison/ (Accessed: 28 June 2025).

[2] AltexSoft. *Working with the Pandas Library: Pros and Cons*. Available at: https://www.altexsoft.com/blog/pandas-library/ (Accessed: 28 June 2025).

5. Can you give an example of how SQL is used to extract insights before applying machine learning?

Example: Predicting Customer Churn

Objective: Identify customers at risk of churn based on their activity and engagement metrics [1].



SQL Query:

```sql
SELECT customer_id,
       MAX(last_login) AS last_login,
       COUNT(DISTINCT session_id) AS num_sessions,
       AVG(session_duration) AS avg_session_duration,
       SUM(purchase_amount) AS total_purchases
FROM customer_activity
WHERE activity_date >= DATE_SUB(CURDATE(), INTERVAL 90 DAY)
GROUP BY customer_id;
```

Source: ChatGPT (2025)

Explanation:

- **MAX(last_login):** Identifies the most recent login date for each customer, which helps understand recent engagement [2].
- **COUNT(DISTINCT session_id):** Counts the number of unique sessions, indicating user engagement frequency [3].

- **AVG(session_duration):** Calculates the average session duration, reflecting user interest and interaction quality [3].
- **SUM(purchase_amount):** Sums the total amount spent, providing insight into customer value and purchasing behavior [4].

Use Case: This summarized data is then used to train machine learning models to predict which customers are likely to churn, by analyzing patterns in their activity and spending [1]

References:

[1] 33rdsquare.com. (2023). *Beginners Guide for Data Analysis Using SQL*. Retrieved from https://www.33rdsquare.com/beginners-guide-for-data-analysis-using-sql/

[2] Wikipedia. (2023). *Data Mining Extensions*. Retrieved from https://en.wikipedia.org/wiki/Data_Mining_Extensions

[3] Enterprise DNA Blog. (2023). *SQL for Data Science: Essential Techniques for Preprocessing and Analysis*. Retrieved from https://blog.enterprisedna.co/sql-for-data-science-essential-techniques-for-preprocessing-and-analysis/

[4] ChatGPT. (2025). SQL query example for data extraction before machine learning.