# The Role of SQL and Data Modeling in AI

Done by: Tufool Alnajashi

AI - Operations

# Why Learn Data Modeling and SQL in AI

## Introduction

Data Modeling and SQL (Structured Query Language) are foundational skills for anyone pursuing a career in Data Science or Artificial Intelligence (AI). AI systems rely on massive amounts of structured, high-quality data. Without the ability to design, store, retrieve, and query data effectively, even the most advanced AI models cannot perform well. This report explores how data modeling and SQL play a critical role in the AI/ML lifecycle and real-world systems.

## Key Insights from Research

### 1. Impact on AI/ML Model Performance

- Data retrieval speed directly influences training time. Efficient SQL queries and indexing reduce latency in accessing training data.  
- Well-modeled databases ensure consistency and reduce redundancy, allowing for cleaner datasets which improve ML model accuracy.  
- Example: In a Google Cloud case study, they noted how clean, structured data pipelines improved the speed and accuracy of ML training for retail demand forecasting.

### 2. Reducing Technical Debt in ML Systems

- Poorly designed schemas and unclean data lead to technical debt: extra future work caused by choosing easy solutions now instead of better approaches.  
- Clean schema design avoids repeated cleaning/transformation steps in production pipelines.  
- Example: According to the paper “Hidden Technical Debt in Machine Learning Systems” (Google researchers), data-related issues like pipeline jungles and undeclared data dependencies are major sources of technical debt.

### 3. Importance for Governance and Monitoring

- SQL databases enable auditing, logging, and versioning of data, crucial for compliance in sectors like finance or healthcare.  
- Example: JPMorgan Chase uses SQL-based lineage tools to monitor data changes in financial models, ensuring accountability and compliance.

## Real-World Examples

- Netflix uses data modeling to manage massive user data across content recommendations and viewership patterns.  
- Airbnb uses SQL-based analytics tools (like Apache Superset) to allow teams to query structured databases and make data-driven decisions.  
- Facebook uses advanced data modeling techniques to manage its vast social graph—representing users, their connections, and interactions. SQL and graph-based queries allow them to personalize feeds, target ads, and detect fake accounts. Their internal tool, **PrestoDB**, is a distributed SQL query engine designed to run fast analytical queries on large datasets.

## Reflection

Through this research, I have recognized that SQL and data modeling are not just database skills—they are critical tools for AI practitioners. In our course exercises, building normalized tables and writing queries prepared us to think logically and cleanly about data. This discipline ensures that data pipelines are efficient, models are trained on trustworthy data, and systems are scalable. Mastering SQL will help me avoid common issues like inconsistent datasets or inefficient queries in real projects.

## References 1. GeeksforGeeks. (n.d.). *SQL for machine learning*. From, <https://www.geeksforgeeks.org/machine-learning/sql-for-machine-learning/>

2.Xebia. (2024, June). *The ever-changing role of data modeling*. <https://xebia.com/blog/the-ever-changing-role-of-data-modeling/>

3. Google Cloud. (2022, October 5). *Making the machine: The machine learning lifecycle*. Google Cloud. <https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle>

4. Presto (SQL query engine) on Wikipedia covers its origins, architecture, and usage at Meta and Netflix, from [Presto (SQL query engine) - Wikipedia](https://en.wikipedia.org/wiki/Presto_%28SQL_query_engine%29?utm_source=chatgpt.com)

5. Facebook (Meta) - *Presto: Facebook’s SQL-on-Anything Query Engineering*, from [Presto: Interacting with petabytes of data at Facebook - Engineering at Meta](https://engineering.fb.com/2013/11/06/core-infra/presto-interacting-with-petabytes-of-data-at-facebook/?utm_source=chatgpt.com)  
Facebook built Presto, a distributed SQL engine, to allow fast querying of large structured data sets across their social network.

e6. ChatGPT (2025). Real-world examples of SQL and data modeling use cases in companies like Facebook and Spotify. OpenAI. Retrieved June 26, 2025, from <https://chat.openai.com>

5. W3Schools SQL tutorials, from

W3Schools. (n.d.). *SQL tutorial*. W3Schools. <https://www.w3schools.com/sql/>