## How Data Storage and Retrieval Affects AI/ML Model Training Performance?

1. **Storage Speed Matters**

   - Slow storage (HDDs or network drives) can bottleneck training
   - Fast storage (like SSDs or in-memory) helps feed data quickly to the model

2. **File Format Impacts Load Time**

   - Text formats (CSV, JSON) are slower to read
   - Binary formats (Parquet, TFRecord, HDF5) are faster and more efficient

3. **Real-Time Preprocessing Can Slow Training**

   - Doing tasks like image resizing or normalization during training increases CPU/RAM load
   - Preprocessing data in advance or using optimized data loaders like (TensorFlow) or (PyTorch) helps to improve performance

4. **Data Shuffling Affects Randomness and Speed**

   - Random access is slower on disk
   - Use shuffle buffers or pre-shuffled datasets for efficient training

5. **Batch Size Is Limited by Storage Efficiency**

   - Inefficient storage means smaller batch sizes, which slow convergence
   - Better data access allows for larger batches using kfold to divide bigger batch size to smaller batches this will lead to faster training

## How does clean, well-modeled data reduce technical debt in production ML systems?

- Making models easier to maintain and debug
- Preventing errors caused by inconsistent or missing data
- Reducing the need for complex code to fix or adjust bad data
- Improving model accuracy and reliability
- Saving time in future updates or retraining

Can you find examples of data governance, monitoring, or auditing that depend on structured databases?

**Data Governance**

- **Access control tables** track who can view or edit data
- **Data dictionaries** stored in relational databases define schema and rules

---

**Monitoring**

- **System logs** stored in SQL tables track data pipeline performance (e.g., load times, errors)
- **Health check dashboards** use structured data to report model status and data freshness

REFERENCE:


1- https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf
2- https://www.tensorflow.org/guide/data_performance
3- https://learn.microsoft.com/en-us/purview/