Deep Lake
A Lakehouse for Deep Learning

Abhinav Tuli
Levon Ghukasyan
Sasun Hambardzumyan

* Image generated by AI
Deep Learning is growing at an unprecedented pace
The Data Problem

projects ends up on the ML project graveyard because of poor data development practices

Source: Rackspace Technologies report
## Solving the problem using data lakes

### Benefits

- Break down Data Silos
- Enable data-driven Decision Making
- Improve Operational Efficiency
- Reduce Costs

### Limitations

- Complex data isn’t supported
- No Deep Learning integration
- Missing gap between MLOps and MDS
- Queries only for analytics
Introducing Deep Lake: Lakehouse for Deep Learning

1. Tensor Data Format
2. Version Control
3. ML Focused Queries
4. Visualisation
5. Streaming
Tensor Storage Format: Native to Deep Learning
Version Control: Track Data Lineage
**Query**: Rapid queries with Tensor Query Language (TQL)

```
SELECT images [100:500, 100:500], boxes + ARRAY[-100, -100, 0, 0]
WHERE contains(categories, 'bicycle') and weather == 'raining'
ORDER BY AOI(boxes, prediction) desc
LIMIT 1000
```

Example query with indexing tensors inline with select and ordered by user-defined function computation.
Visualize: In-browser visualization engine
Visualize: In-browser visualization engine
Stream: Streaming Data Loaders up to 4x faster*

*Ofeidis et al; An Overview of the Dataloader Landscape. Challenges and Promises
Save Time: Access LAION Dataset in <5 Seconds

100 hrs+

with in-house solutions

Before: no indexing and shuffling with WebDataset

After with DeepLake: + upload in ~6 hr

5 seconds

with activeloop

Download Parquet 50GB Files

Fetch the data file by file

Shuffle, Crop and convert to WebDataset format

Finally load the data

Access a particular slice of the data (on average)

Stream any slice of the data as if it were on your PC

import deeplake
ds = deeplake.load('hub://laion/laion-400M')
dsv = ds.query('select * where ...')
dl = dsv.pytorch(num_workers=16)
Deep Lake: Lakehouse for Deep Learning

Machine Learning Loop with Deep Lake
Unlock compute from data bottleneck
Dive into Deeplake
https://github.com/activeloopai/deeplake