RAPIDS: end-to-end data science

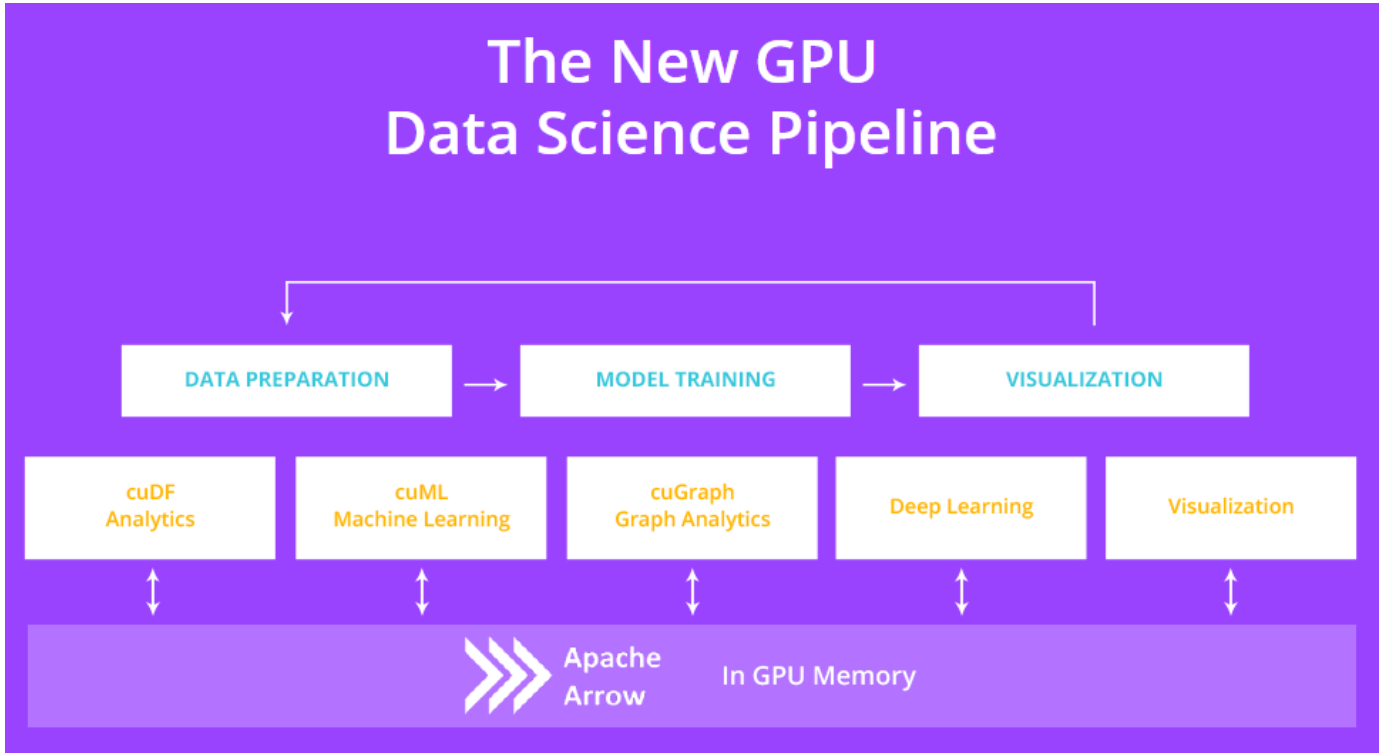
Michael Poli

RAPIDS in a nutshell:

RAPIDS is a suite of open source software libraries aim to enable execution of end-to-end data science and analytics pipelines entirely on GPUs.

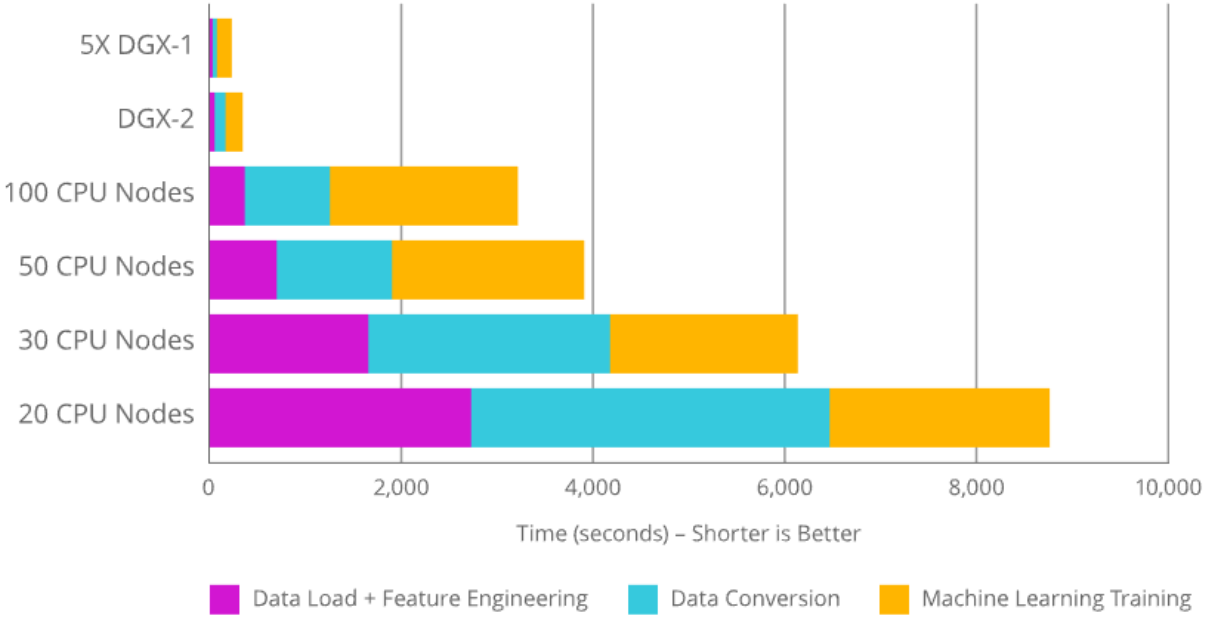
These libraries rely on NVIDIA® CUDA® primitives for low-level compute optimization, while concurrently exposing GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces.

https://rapids.ai/index.html



https://rapids.ai/index.html

End-to-End Faster Speeds on RAPIDS



https://rapids.ai/index.html

KEY FEATURES:

cuDF:

DataFrame manipulation for data preparation

cuDF currently has 2 APIs: C++ and a Python API that mimics Pandas

cuML:

GPU-accelerated machine learning libraries.

Apache Arrow:

Columnar in-memory data structure that delivers efficient and fast data interchange with flexibility to support complex data models.

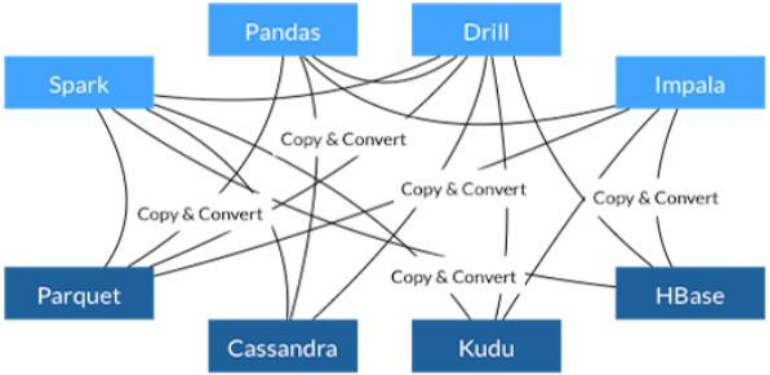
https://rapids.ai/index.html

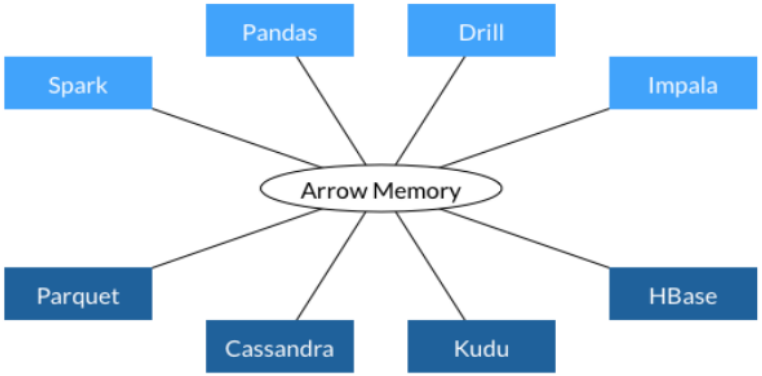
Apache Arrow

Apache Arrow is a cross-language development platform for in-memory data.

Standard language-independent columnar memory format for flat and hierarchical data, organized for efficient analytic operations on modern hardware.

https://arrow.apache.org/





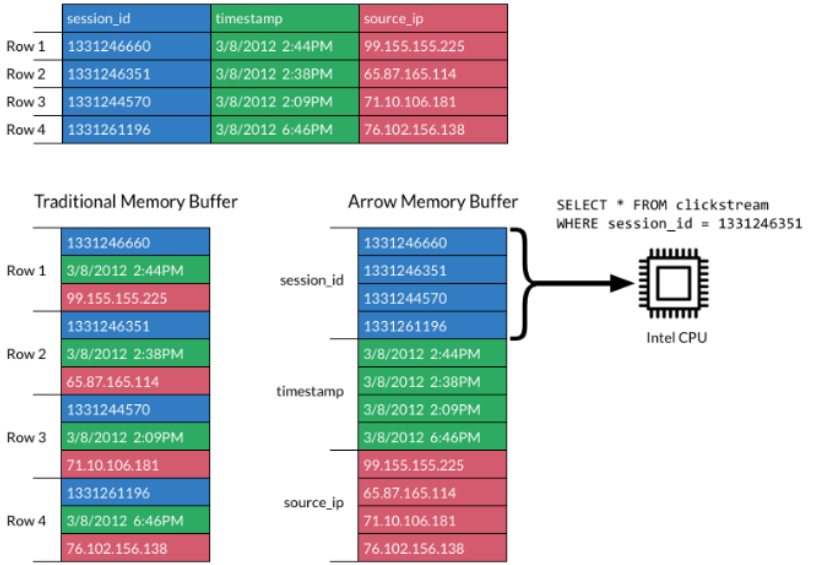
Apache Arrow is an open standard for DataFrame-like data manipulation

Main draw: **zero-copy** reads between interfaces

https://arrow.apache.org/

Additional advantages:

- Execution engines can take advantage of SIMD while processing DataFrames
- Cache misses minimized
- Leverages columnar compression techniques
- Allows for the deconstruction of traditionally vertically integrated analytic database architectures



Apache Arrow: A Cross-Language Development Platform for in Memory Data | SciPy 2018 | Wes McKinney

Example 1:

```
import cudf
gdf = cudf.read_csv('path/to/file.csv')
for column in gdf.columns:
    print(gdf[column].mean())
```

https://rapids.ai/start.html

Example 2: Kaggle competition