Yandex

CatBoost

APACHE Spark™

CatBoost for Apache Spark Architecture

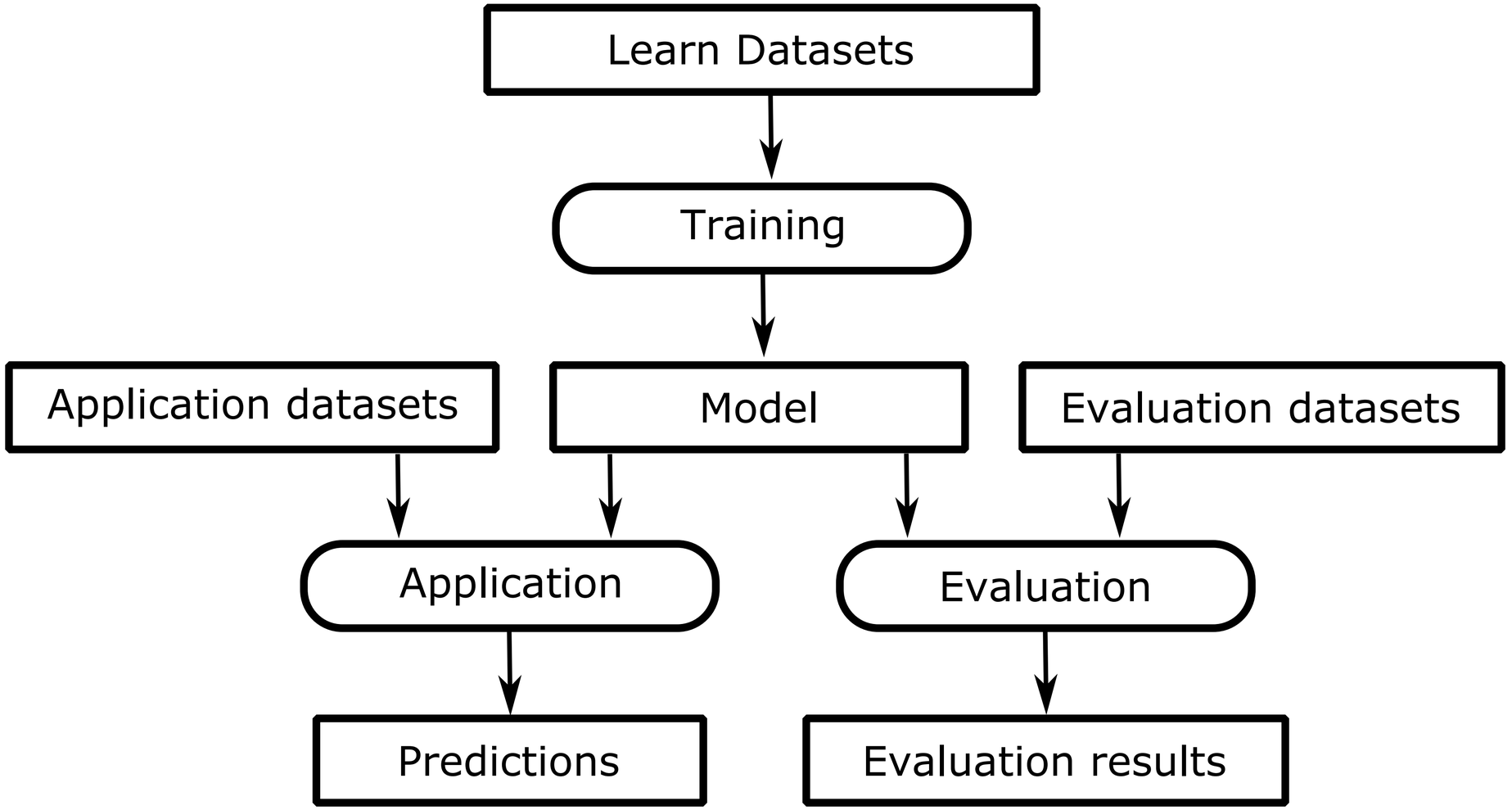
Andrei Khropov,
CatBoost team @ Yandex

What is CatBoost?

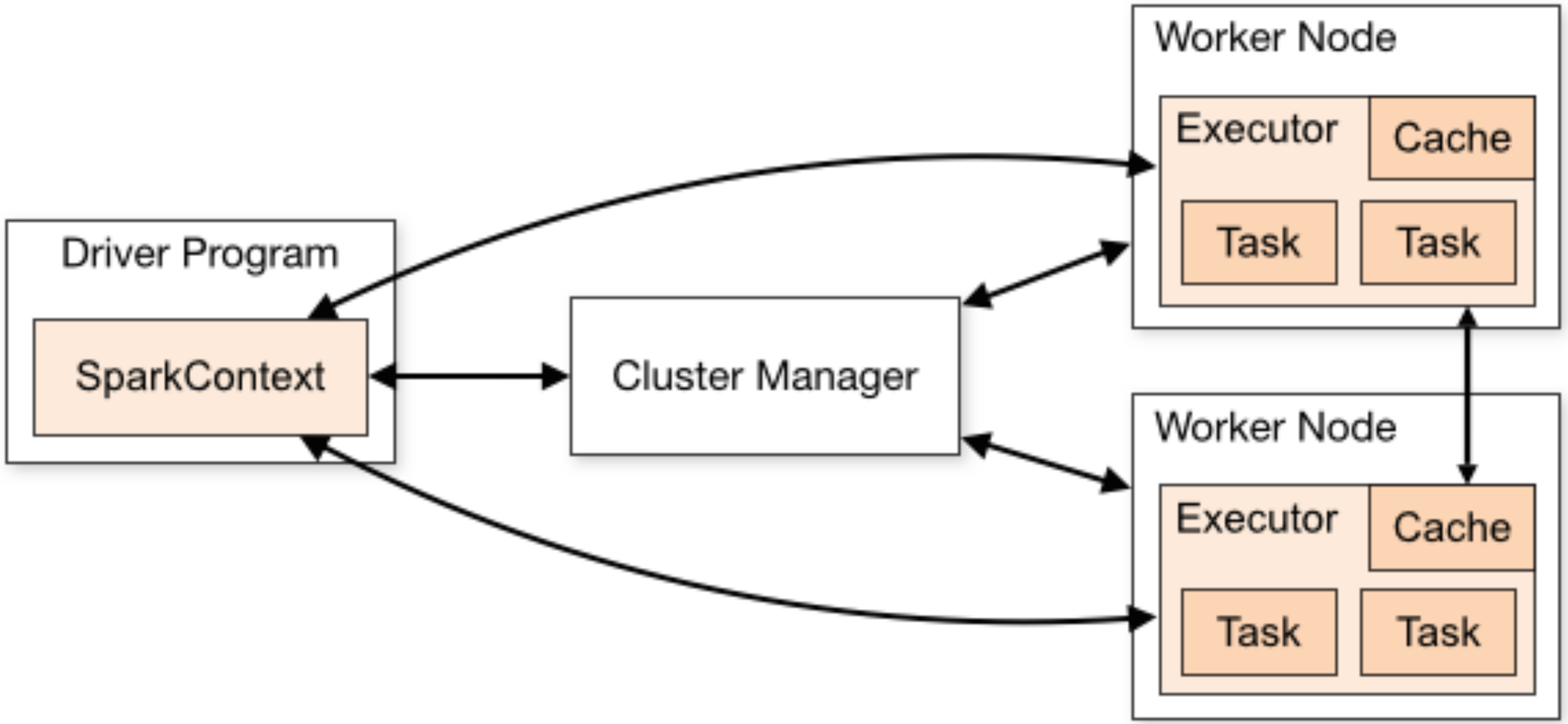
CatBoost

- CatBoost is a machine learning algorithm that uses gradient boosting on decision trees (GBDT).
- It is available as an open source library.
- See “Introducing CatBoost for Apache Spark” presentation for more basic info.

Generic Machine Learning Workflow

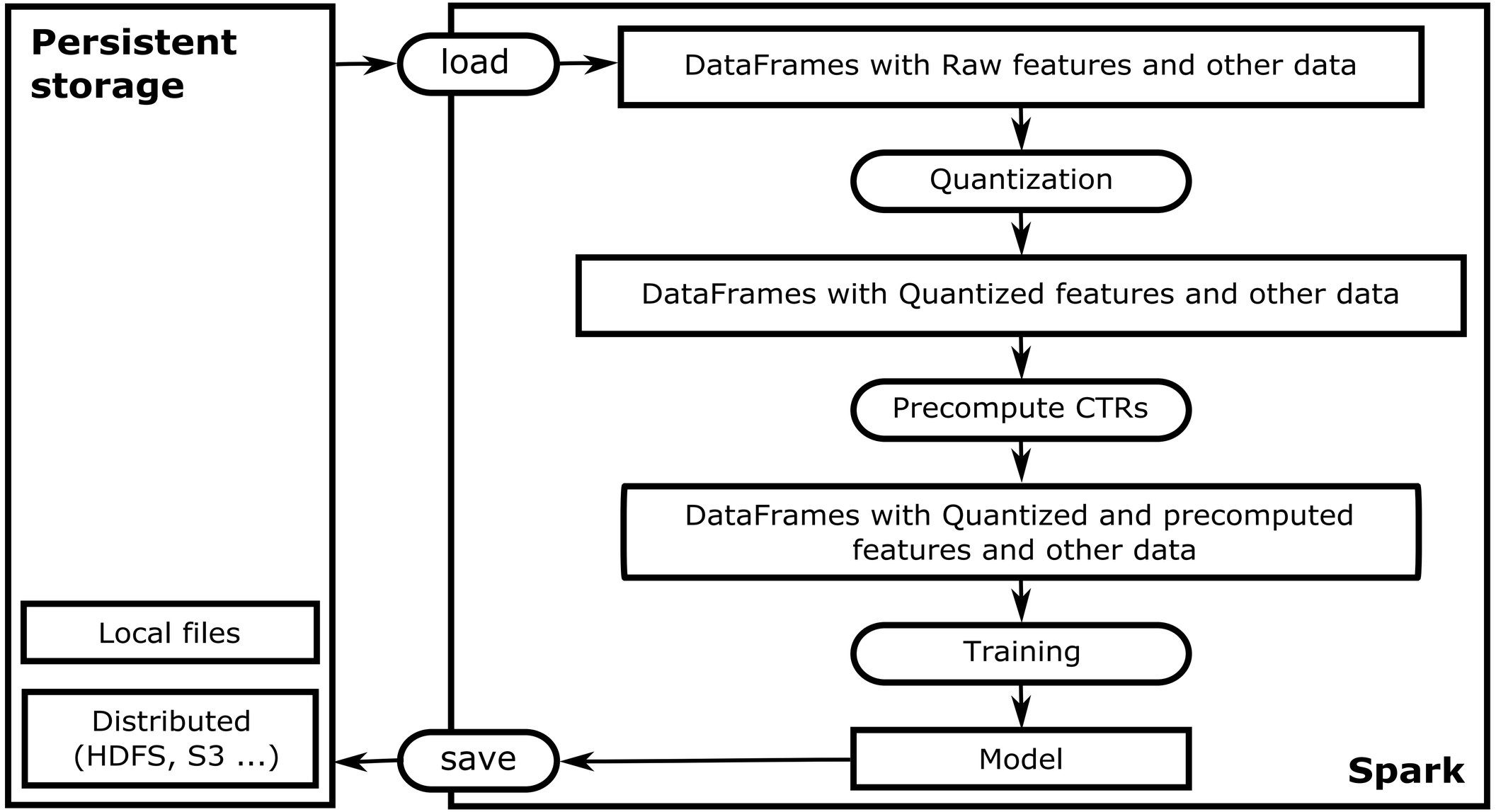


Spark Cluster Overview

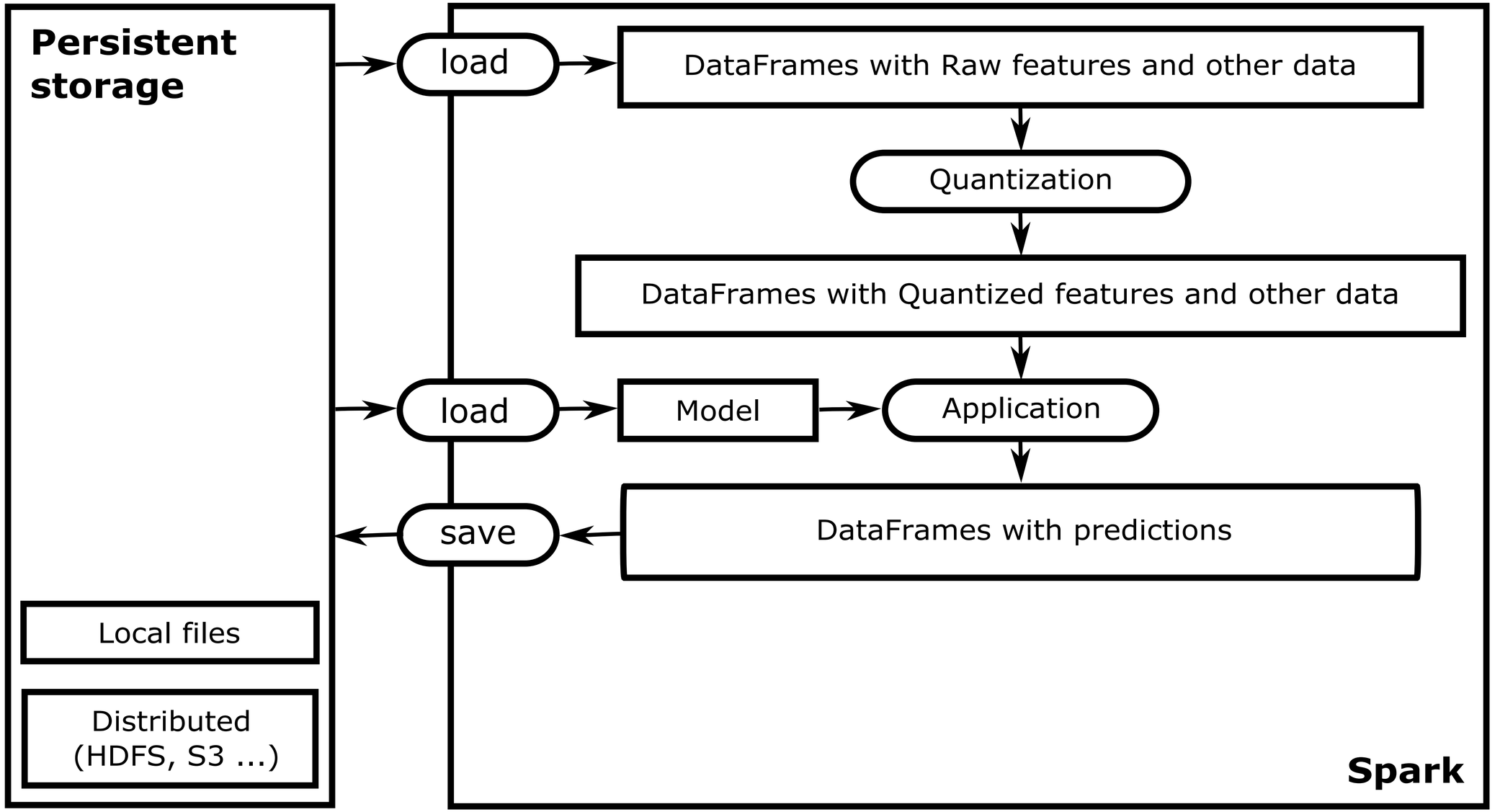


› Image credits: https://spark.apache.org/docs/latest/cluster-overview.html

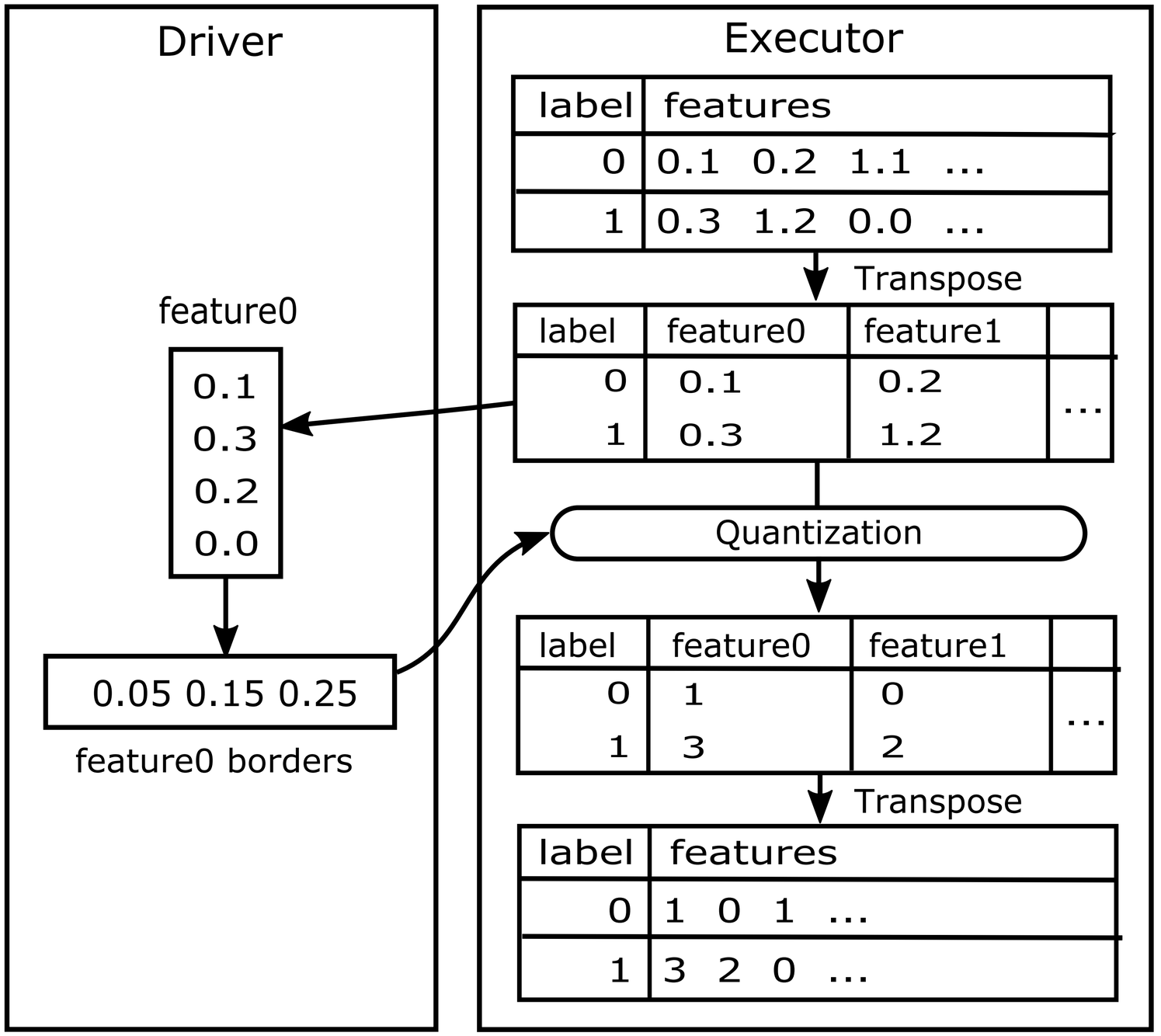
Training Workflow



Application Workflow

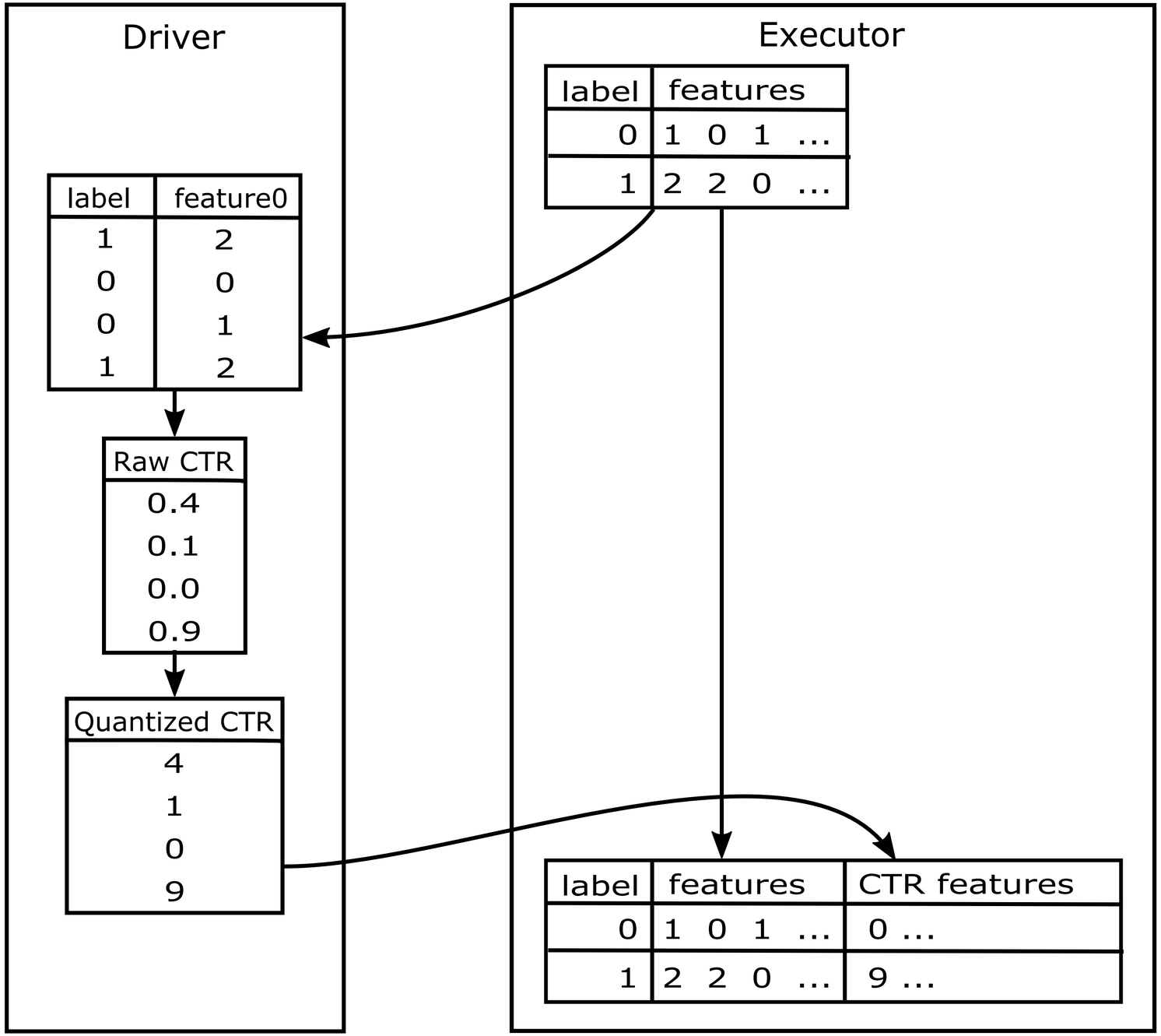


Quantization



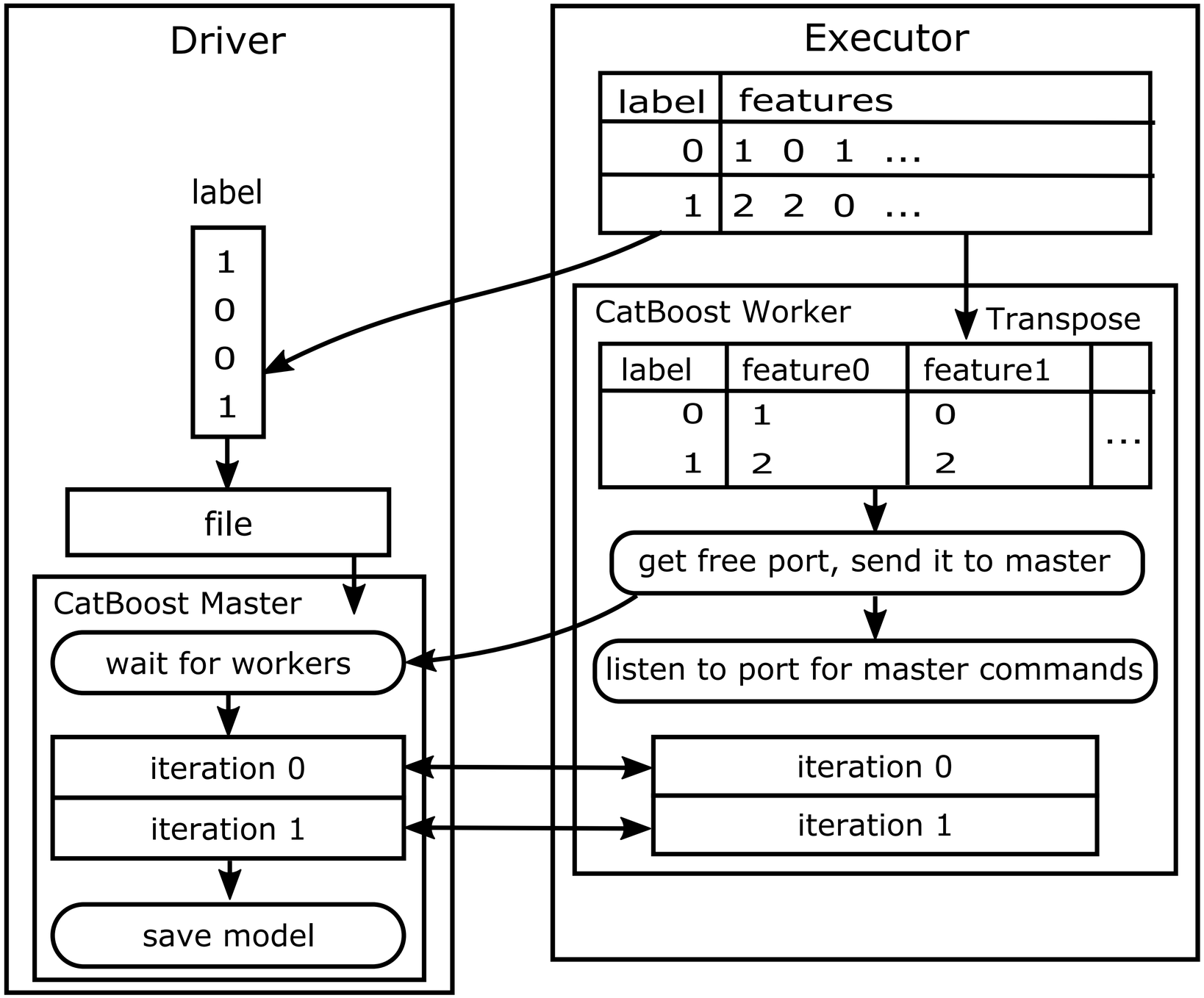
CatBoost APACHE Spark™

CTRs Precalculation



CatBoost APACHE Spark™

Training



CatBoost APACHE Spark™

The End

CatBoost

- CatBoost website: https://catboost.ai/
- CatBoost documentation: https://catboost.ai/docs
- CatBoost on GitHub: https://github.com/catboost
- CatBoost for Apache Spark home: https://github.com/catboost/catboost/tree/master/catboost/spark/catboost4j-spark