

CatBoost



# Как запихнуть в CatBoost терабайты данных используя Apache Spark

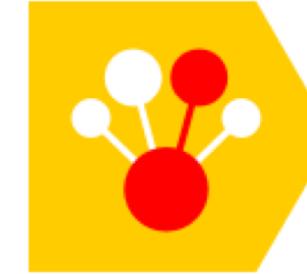
Андрей Хропов,  
Старший разработчик в Яндексе

# О чём воркшоп?



- › Сравнение с конкурентами
- › Архитектура CatBoost для Apache Spark для понимания
- › Практические аспекты запуска обучения CatBoost на Apache Spark
- › Примеры кода (на PySpark, на Java/Scala - аналогично)
- › Примеры будут не на терабайтах – так быстрее
- › Введение в CatBoost для Apache Spark – см. на канале «Разработка» и презентации в репозитории туториалов CatBoost

# CatBoost Spark и конкуренты

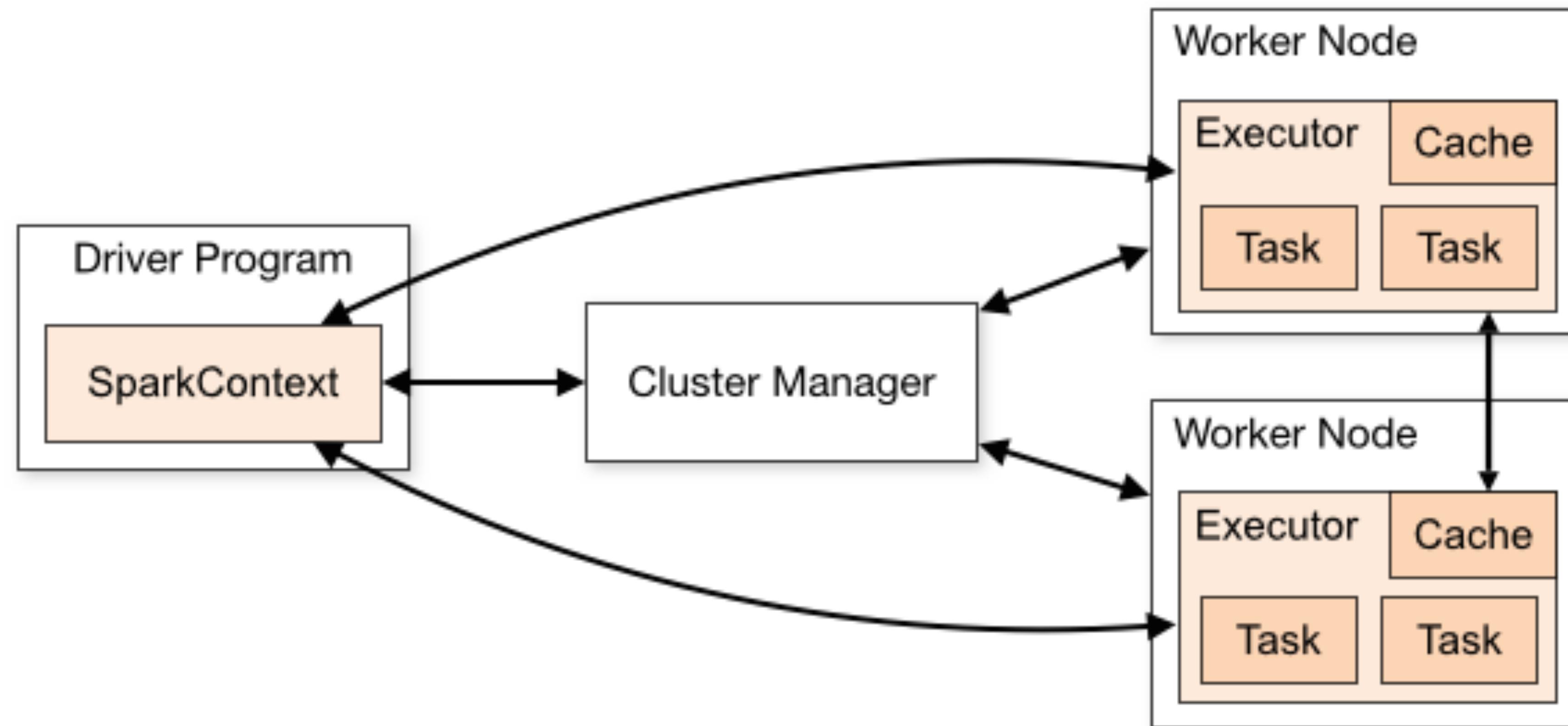
	 CatBoost	 LightGBM	 XGBoost
Поддержка GPU	Нет (планируется)	Нет	Да
Категориальные признаки	Да, включая CTR статистики	Да	Только представленные как one hot
Пары	Да	Нет	Нет
Пред-квантование	Да	Нет	Нет
Поддержка PySpark	Да	Да	Нет
Поддержка SparkR	Нет	Бета	Нет

# Производительность CatBoost Spark и конкурентов

	Criteo derived 170 м примеров 65 признаков		Epsilon 400 к примеров 2000 признаков		Higgs 10,5 м примеров 28 признаков	
	всего	на итерацию	всего	на итерацию	всего	на итерацию
 <b>CatBoost</b>	53 м 52 с	2,8 с	1 ч 5 м	3,7 с	7 м 40 с	0,31 с
 <b>LightGBM</b>	59 м	3,5 с	2 ч 36 м	9,4 с	2 ч 25 м	8,7 с
 <b>XGBoost</b>	около 8 ч	28,9 с	17 м 25 с	1 с	10 м	0,6 с

- › Конфигурация кластера – 16 машин x 16 ядер
- › Время на 1000 итераций, общее время включает предобработку

# Кластер Spark

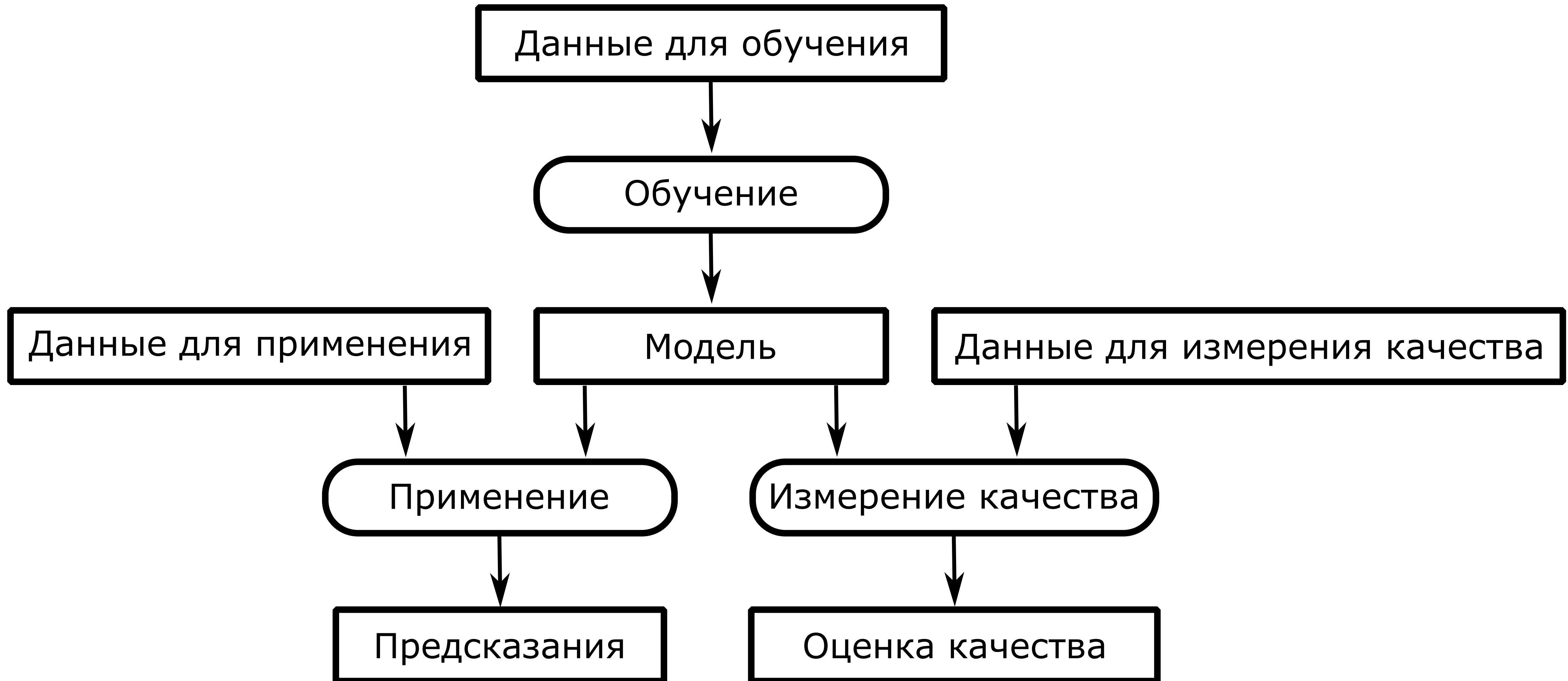


# Кластер Spark

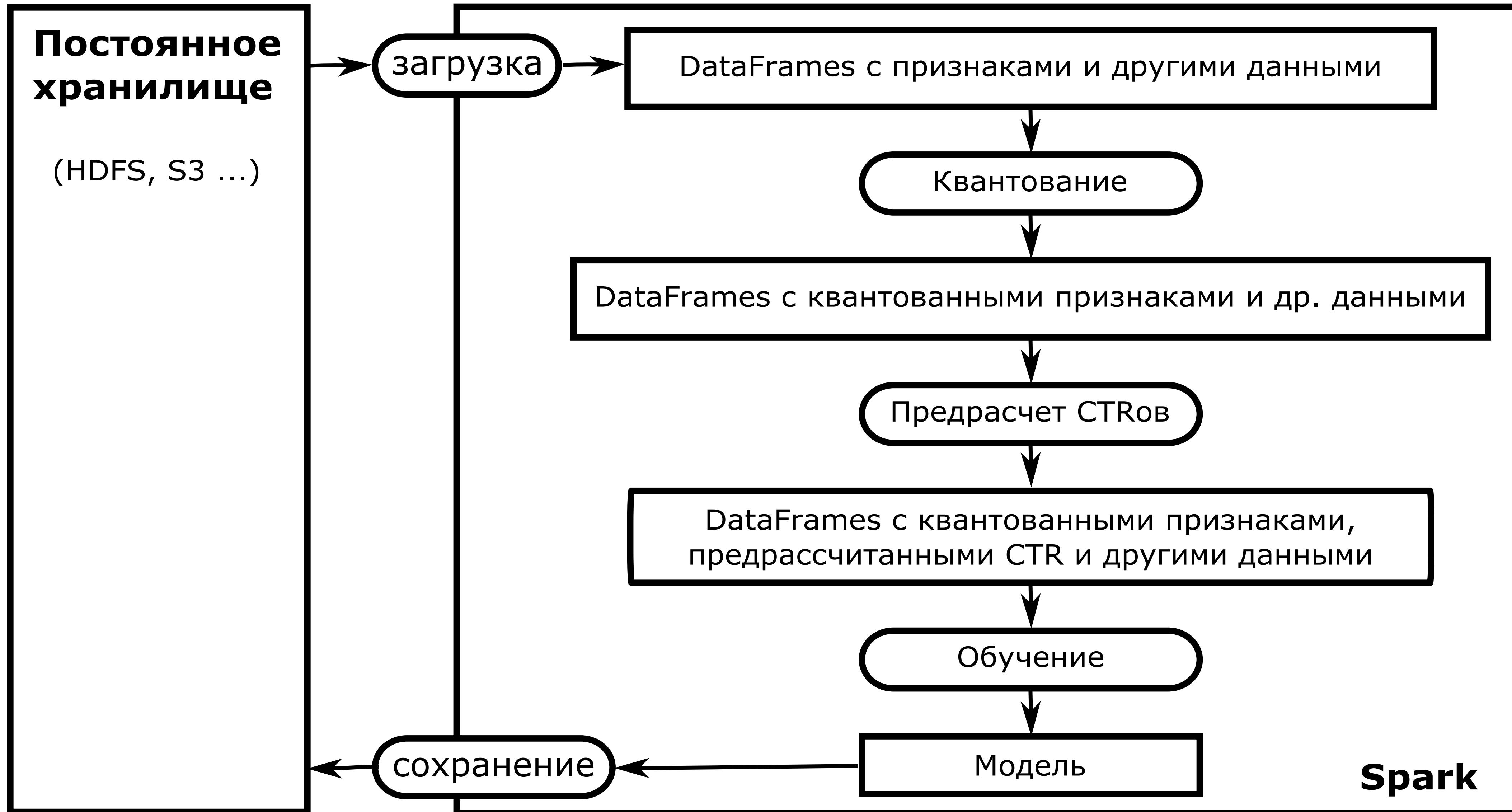


- › Локальная машина
  - Разработка, тесты, CI
- › Свой
  - + Максимальный контроль
  - Сложность масштабирования
  - Большие расходы сразу
  - Потенциальная неэффективность (простой)
- › Облачный
  - Меньше контроля
  - + Масштабируемость
  - Расходы при 100% загрузки типично выше
  - + Гибкость

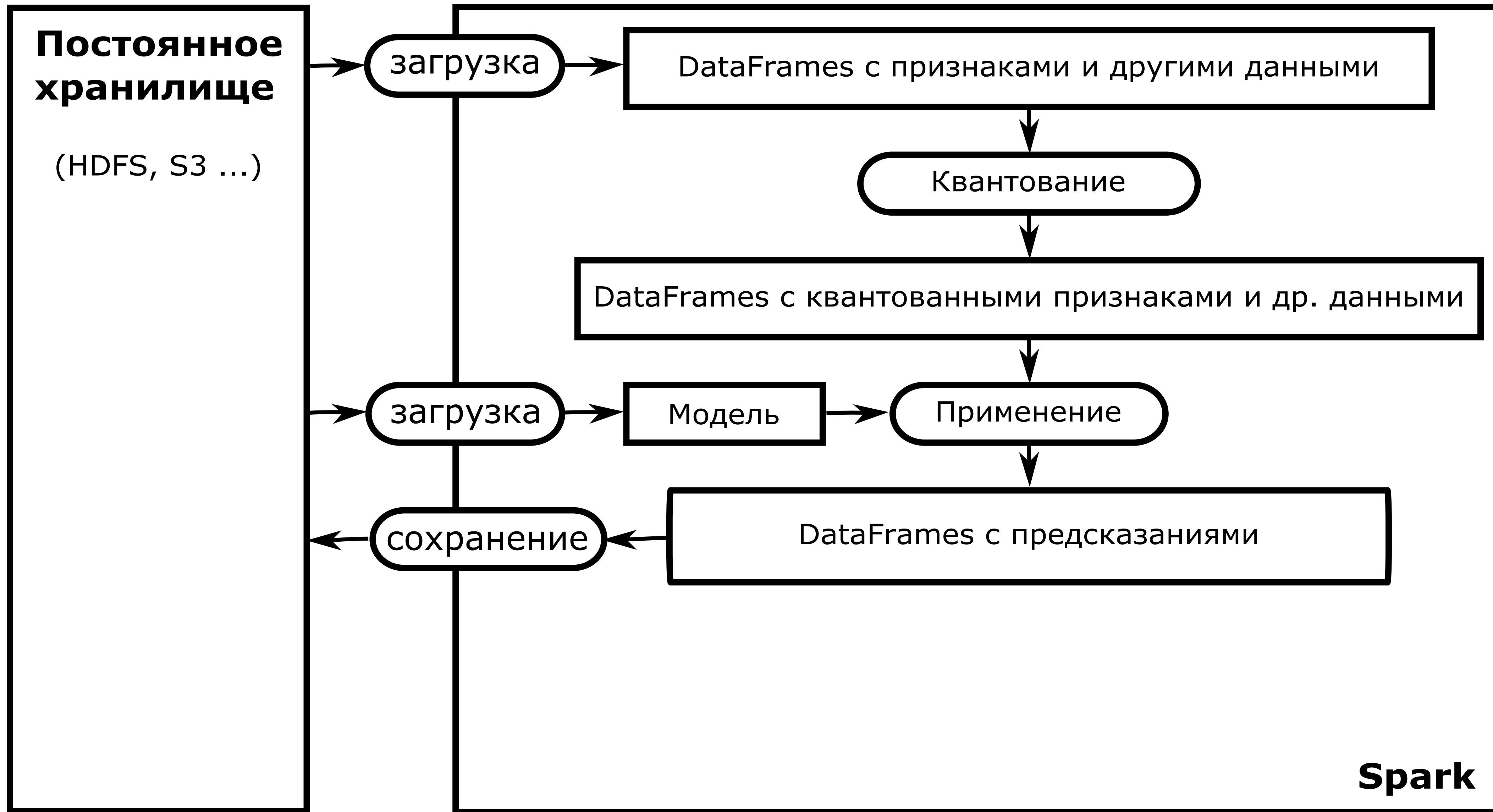
# Общий процесс машинного обучения



# Процесс обучения

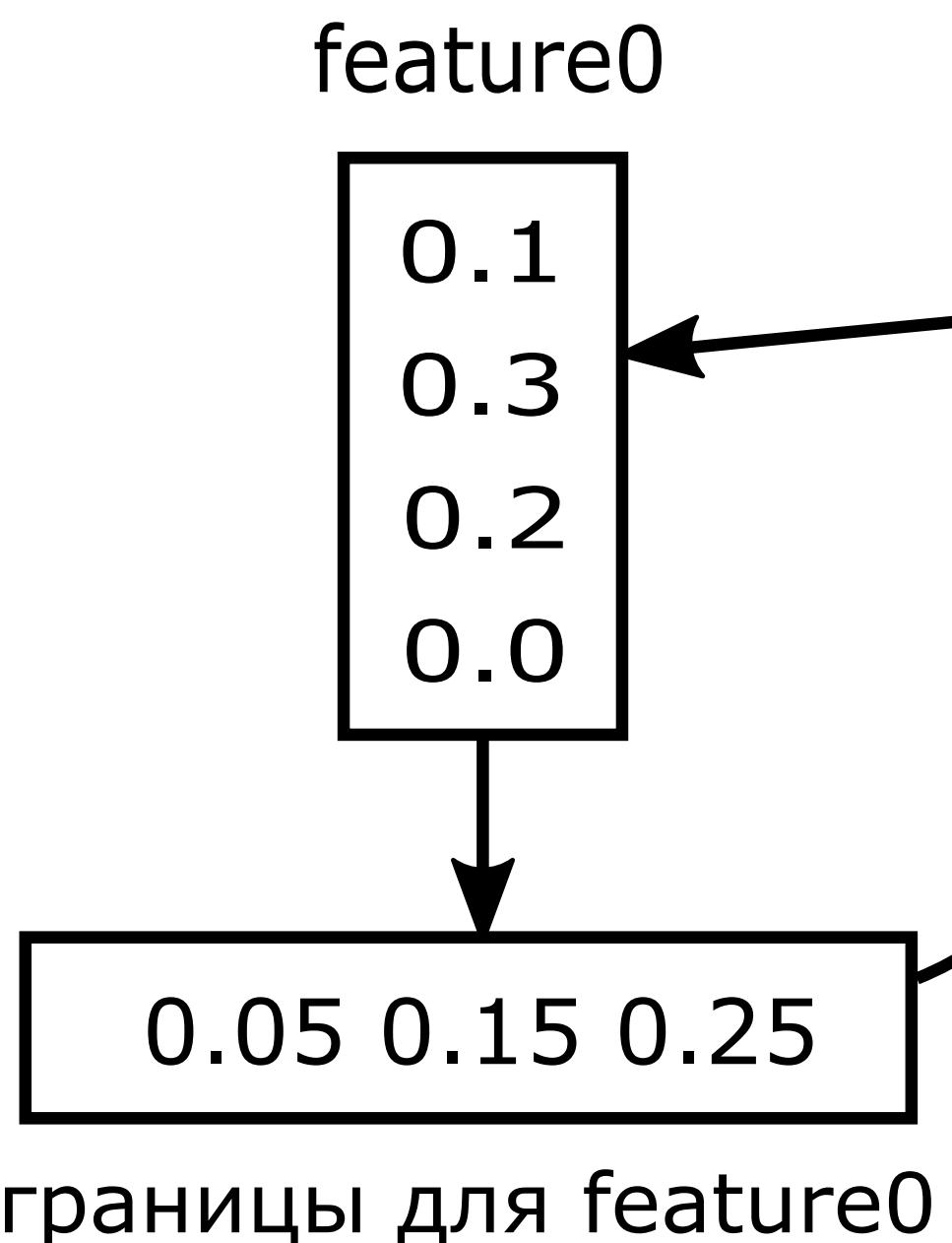


# Процесс применения



# Квантование

Driver



Executor

label	features
0	0.1 0.2 1.1 ...
1	0.3 1.1 0.0 ...

Транспонирование

label	feature0	feature1	
0	0.1	0.2	...
1	0.3	1.1	

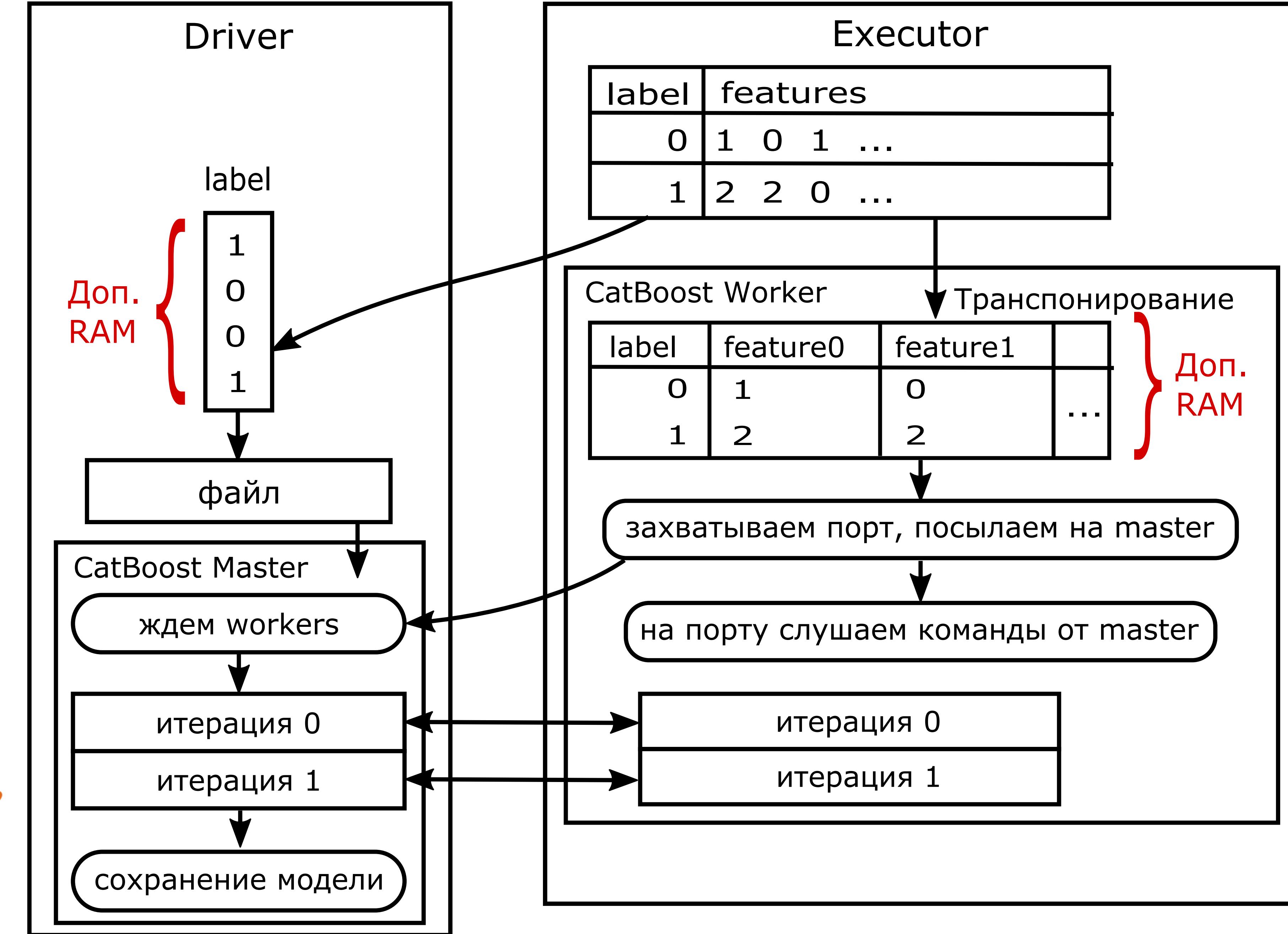
Квантование

label	feature0	feature1	
0	1	0	...
1	3	2	

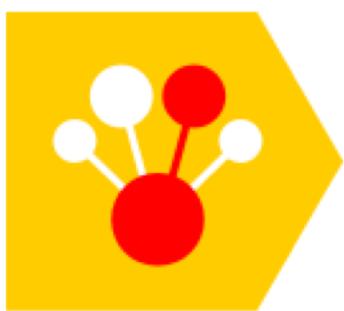
Транспонирование

label	features
0	1 0 1 ...
1	3 2 0 ...

# Обучение



# Вопросы?



CatBoost



- › Сайт CatBoost: <https://catboost.ai/>
- › Документация CatBoost: <https://catboost.ai/docs>
- › CatBoost на GitHub: <https://github.com/catboost>
- › Главная страница CatBoost для Apache Spark:  
<https://github.com/catboost/catboost/tree/master/catboost/spark/catboost4j-spark>