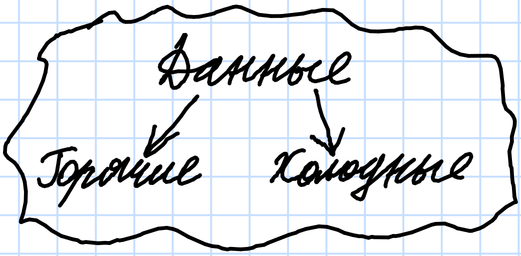
Data Lake могут поддерживать только большие IT компании.

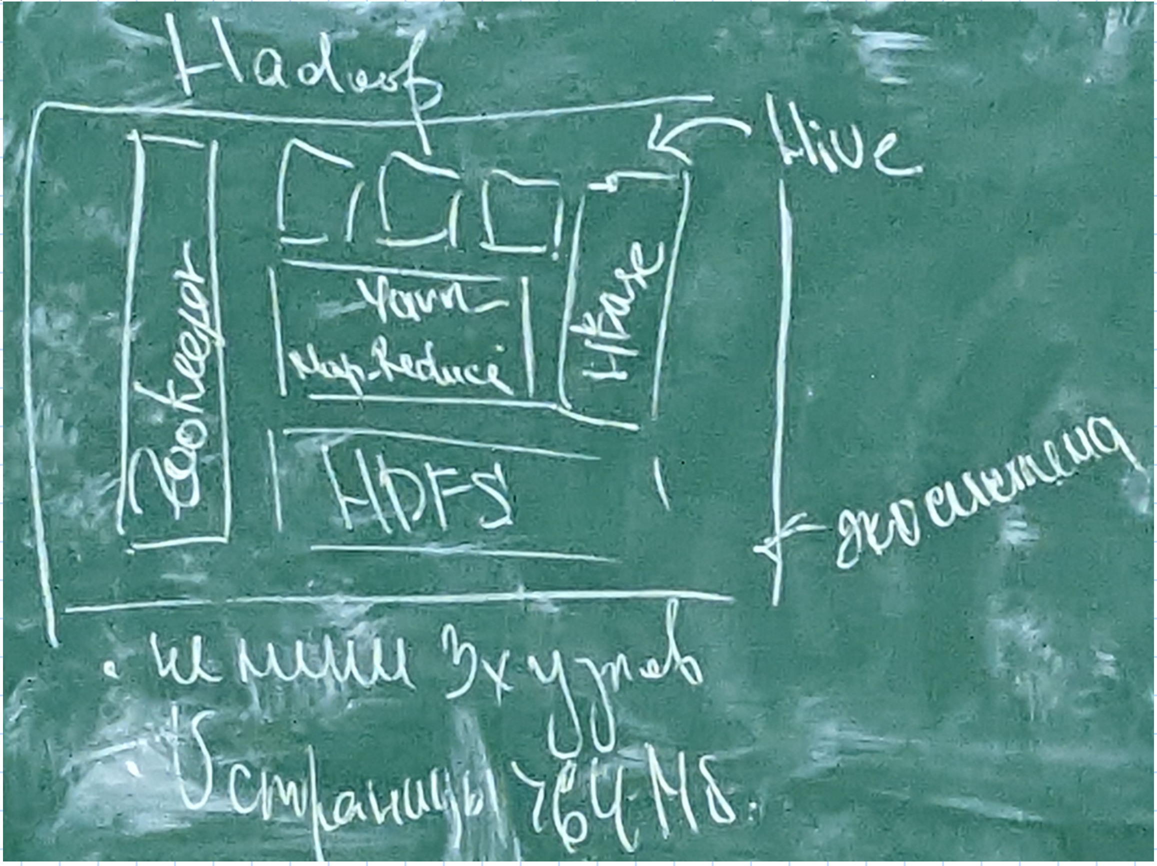
Hadoop — распределённое хранилище данных (написана на Java)

(именно холодных) → не часто нужны, но важно их иметь у себя



Тенденция вынесения холодных данных в отдельные сервера ⇒ Hadoop (ориентировована на большие данные)

Основан на своей файл. системе HDPS (1 страница – 64 Мб, а не 4/8 Кб)



Компоненты универсальны
(могут использоваться отдельно
в других системах)
(HDFS исп. в Clickhouse)

Hadoop: хранение и расчёт холодных данных, фоновые задачи (а ля „запустить раз в сутки/ /месяц/год)

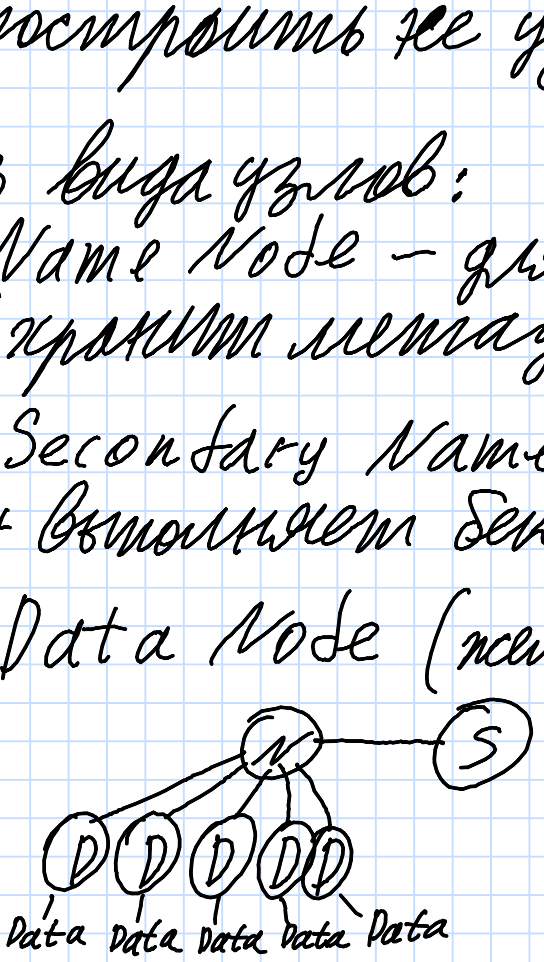
Не Realtime!

В крупных компаниях рядом с Hadoop стоит OLAP-система (для горячих данных): Clickhouse, Vertica, ...

Аналитические системы на Hadoop построить не удалось

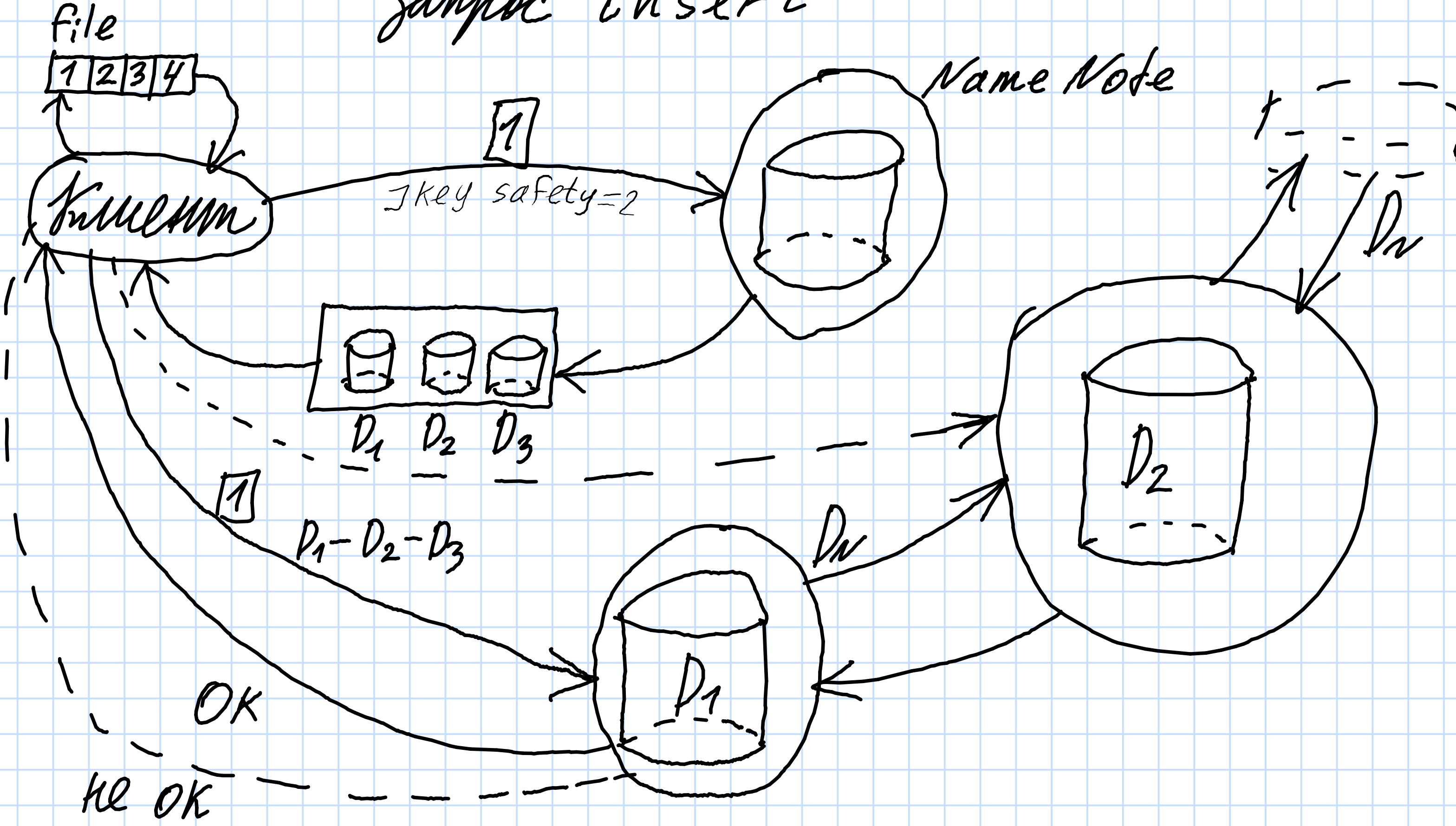
3 вида узлов:

1) Name Node – для маршрутизации;
(хранит метаданные + контакт с клиентом)
2) Secondary Name Node – дублёр Name Node +
+ выполняет бекапы журналов (репликация)
3) Data Node (желательно ≥ 3 шт.)



Типовое взаимодействие:

запрос insert



Узлы сами сигнализируют о своём состоянии;

Алгоритм Map-Reduce.

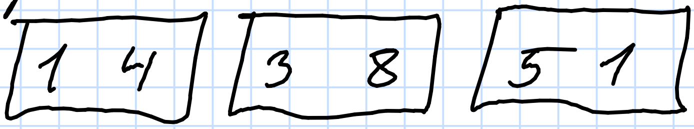
Шаги:

1) Map- распределение задач по серверам; (на потоки)

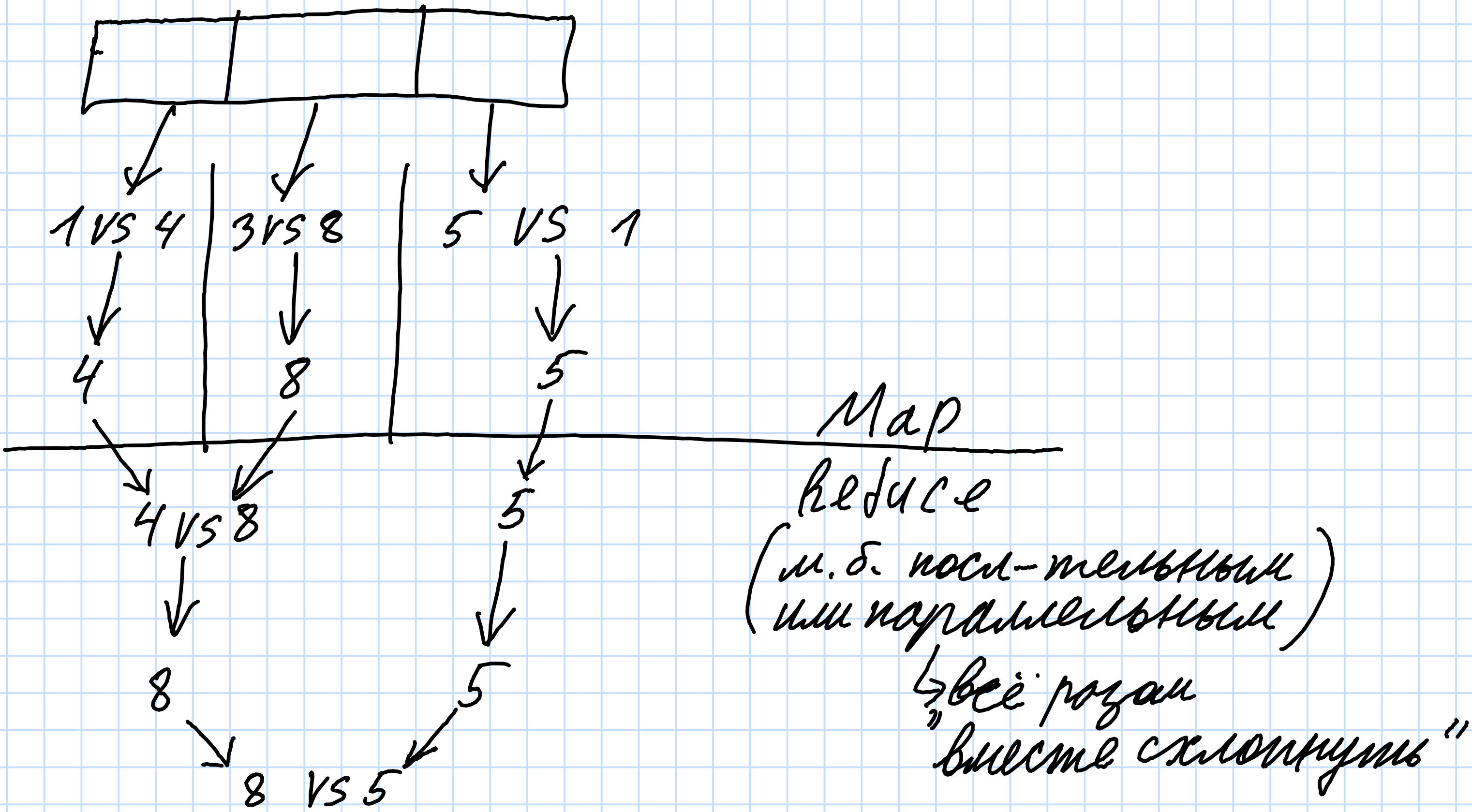
Hadoop не применяется для посл-тельных и рекуррентных задач, т.к. они плохо распараллеливаются;

2) Reduce;

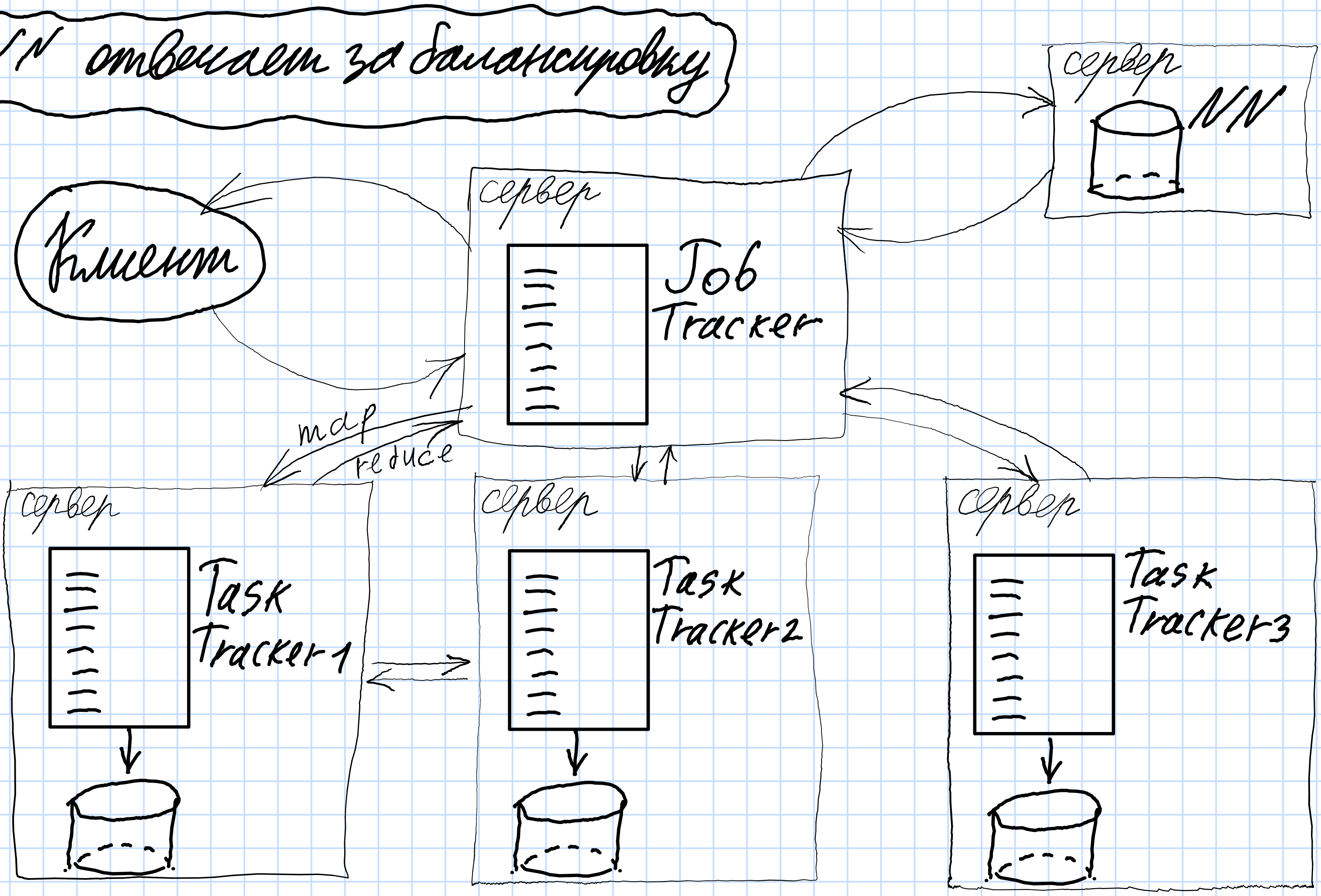
Пример: нахождение максимума в массиве



max = ?



NN отвечает за балансировку


сервер
NN
Клиент
сервер
Job Tracker
map
reduce
сервер
Task Tracker-1
сервер
Task Tracker2
сервер
Task Tracker3