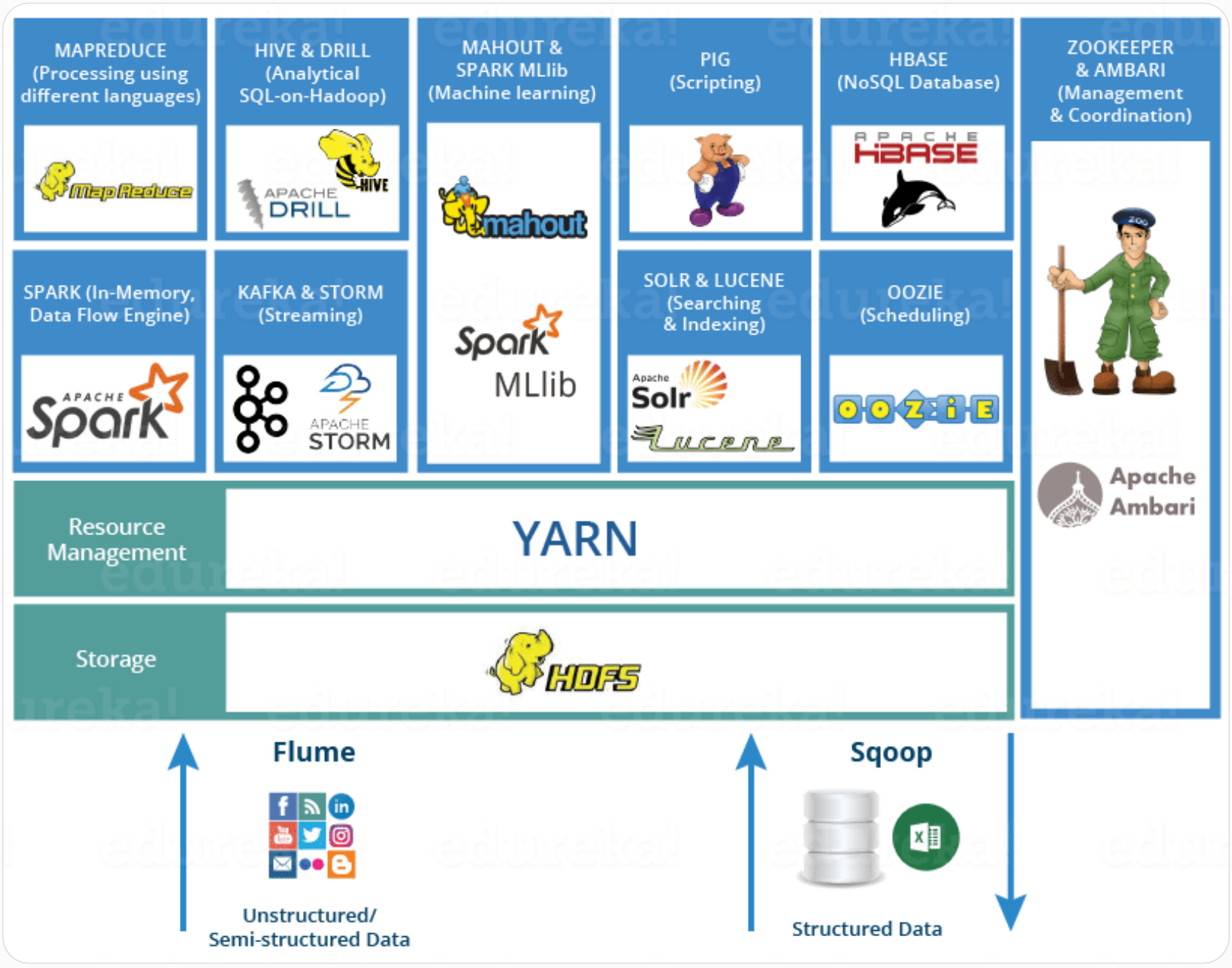
Экосистема Hadoop

Экосистема Hadoop

- Элементы экосистемы Hadoop
- Распределённая файловая система HDFS
- Фреймворк распределённых вычислений MapReduce
- Отличия хранилища данных на Hadoop от БД

Экосистема Hadoop



Изображение: Экосистема Hadoop

Обзор элементов экосистемы Hadoop

Загрузка данных из внешних источников:

структурированные

Sqoop

неструктурированные
и слабоструктурированные

Flume

Хранение данных:

HDFS

Работа с данными:

реляционные БД

Flume

NO SQL

HBASE

Распределённые вычисления:

MapReduce

Spark

Мониторинг ресурсов:

YARN

Мониторинг элементов экосистемы:

Ambary

Схема: Skillbox

Файловая система HDFS

HDFS (Hadoop Distributed File System) — файловая система, предназначенная для хранения файлов больших размеров, поблочно распределённых между узлами вычислительного кластера.

Принципы работы HDFS

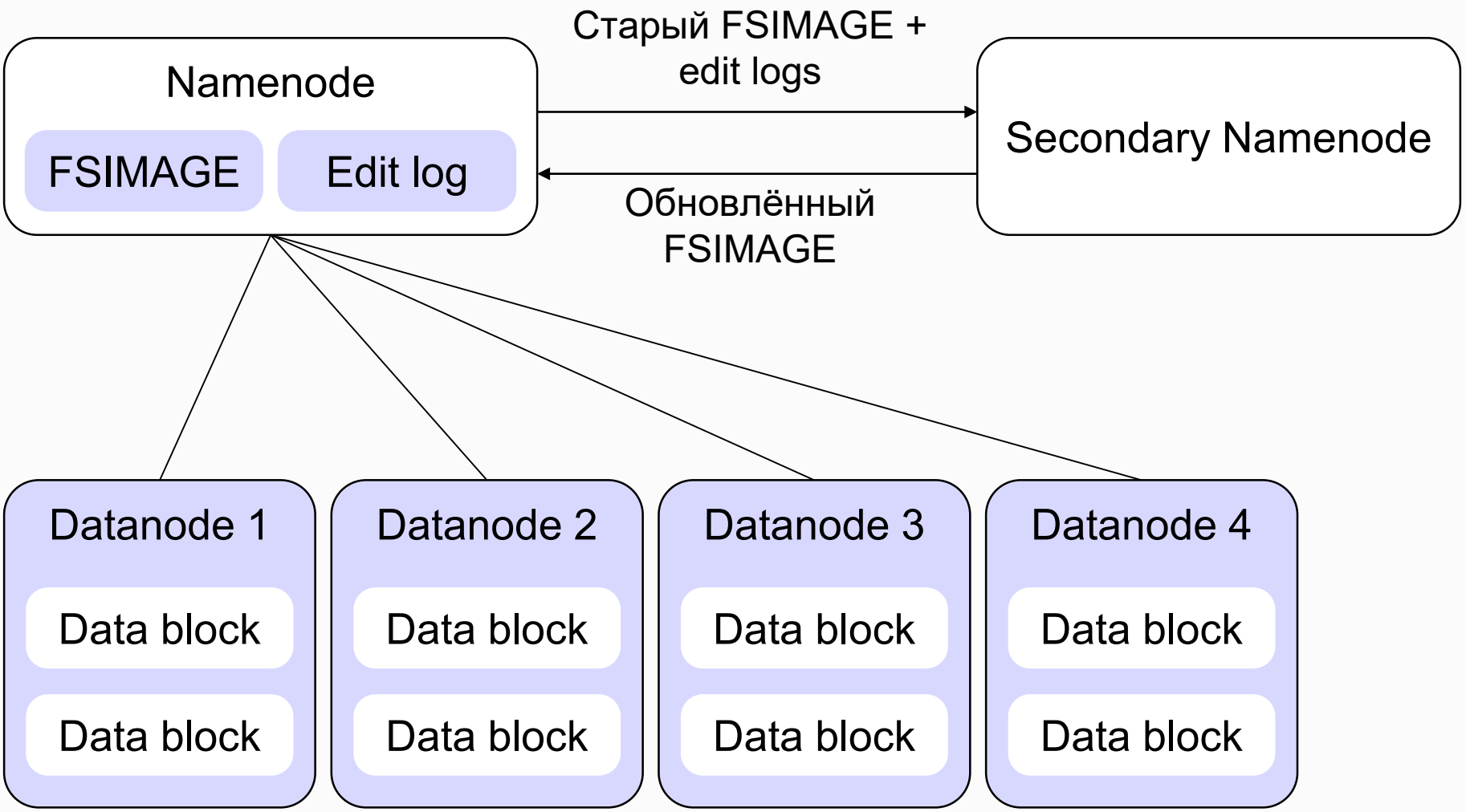


Схема: Skillbox

Принципы работы HDFS

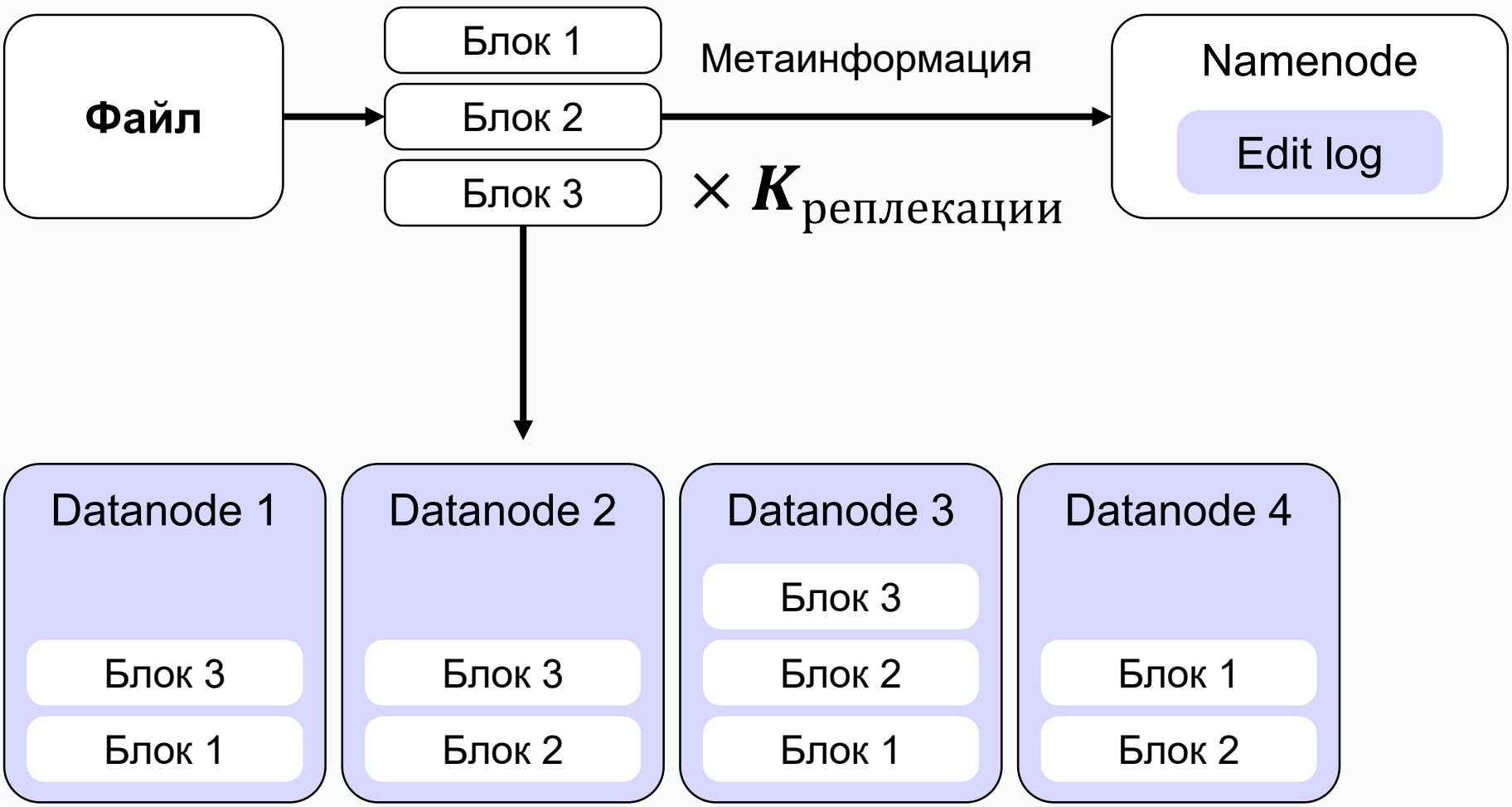


Схема: Skillbox

Принципы работы HDFS

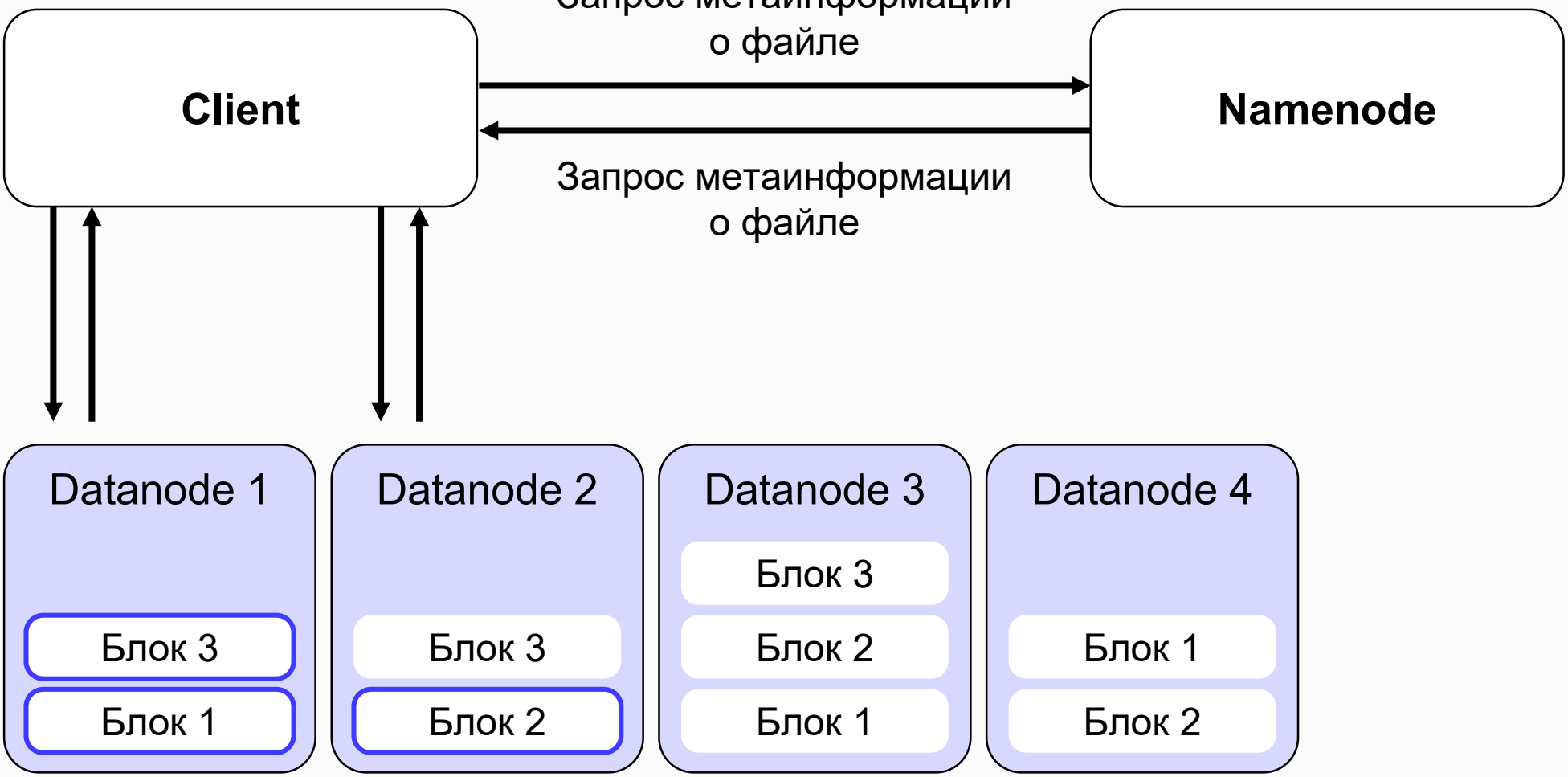


Схема: Skillbox

MapReduce

MapReduce — модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений над очень большими, вплоть до нескольких петабайт, наборами данных в компьютерных кластерах.

Map — трансформация данных.

Пример из SQL:

```
SELECT substring(capital, 1, 1)
as first_letter FROM …
```

Reduce — агрегация данных.

Пример из SQL:

```
… SELECT first_letter, count(*) FROM …
GROUP BY first_letter
```

MapReduce

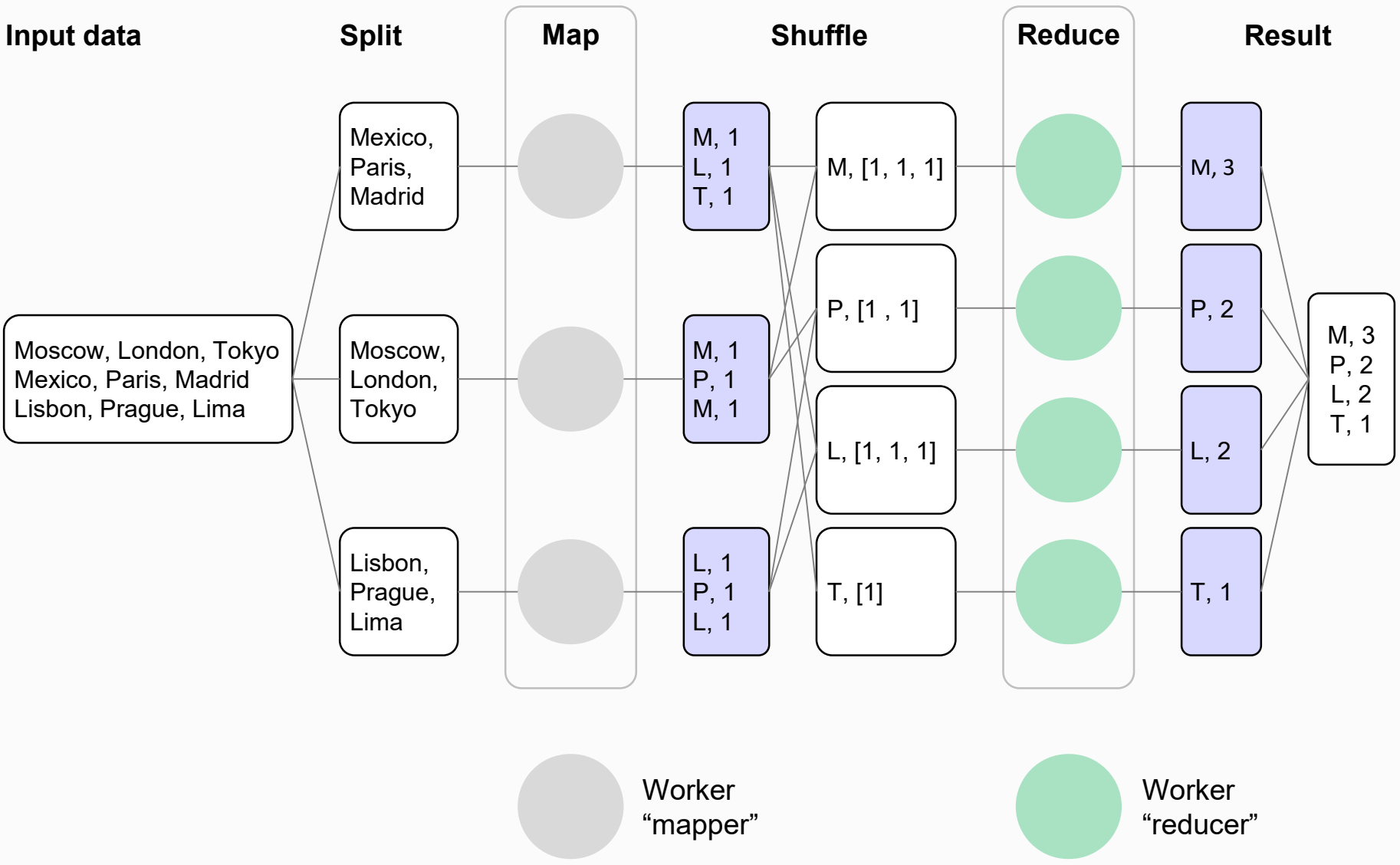


Схема: Skillbox

MapReduce

Особенности

- MapReduce подходит для обработки больших файлов в силу своей горизонтальной масштабируемости, но не подходит для малых файлов
- В силу того, что данные делятся на блоки, которые обрабатываются независимо, MapReduce не сможет работать с архивами или с файлами, блоки которых логически связаны

MapReduce

	Database	Data Warehouse
Структура данных	Структурированные данные, которые могут быть организованы в связанные друг с другом таблицы	Структурированные и неструктурированные данные, включая файлы, изображения, видео и текст
Временные интервалы	Сконцентрированы на текущих или оперативных данных	Как правило содержит данные за длительные периоды (историчность)
Цели применения	Для оперативной обработки транзакций	Отчётность, анализ данных, бизнес-аналитика
Обновление данных	Данные доступны в режиме реального времени	Данные обновляются по мере необходимости (напр., раз в день)

Таблица: Skillbox

Выводы

- Узнали, из каких элементов состоит Hadoop
- Узнали, как в Hadoop хранятся большие данные
- Узнали, как Hadoop обрабатывает большие данные