**MapReduce**

**Что такое MapReduce?**

* модель распределенной обработки данных и исполнительная среда, работающая на больших кластерах типовых машин
* Все запуски map и reduce работают независимо и параллельно
* Map выполняется на той же машине, на которой хранятся данные
* MapReduce – это всегда полное сканирование данных

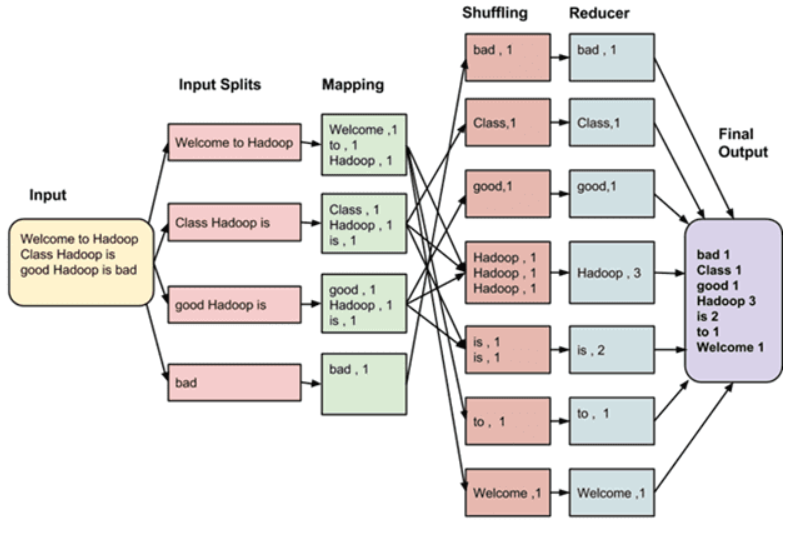
**Преимущества**

* Масштабируемость
* Гибкость
* Скорость обработки
* Простота (Относительная)

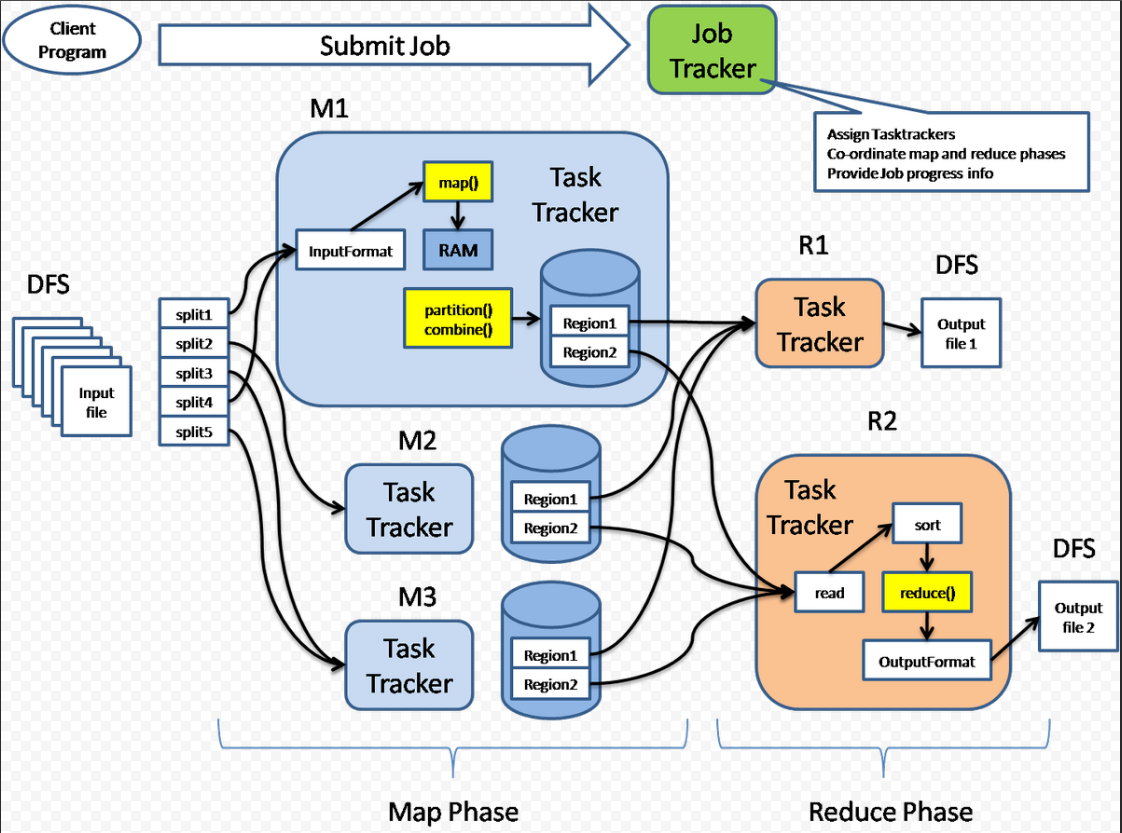
**Недостатки**

* **недостаточно высокая производительность** – классическая технология, в частности, реализованная в ядре Apache Hadoop, обрабатывает данные ациклично в пакетном режиме. При этом функции Reduce не запустятся до завершения всех процессов Map. Все операции проходят по циклу чтение-запись с жесткого диска, что влечет задержки (latency) в обработке информации.
* **ограниченность применения**– высокие задержки распределенных вычислений, приемлемые в пакетном режиме обработки, не позволяют использовать классический MapReduce для потоковой обработки в режиме реального времени, повторяющихся запросов и итеративных алгоритмов на одном и том же датасете, как в задачах машинного обучения ([Machine Learning](https://www.bigdataschool.ru/wiki/machine-learning" \t "_blank)). Для решения этой проблемы, свойственной Apache Hadoop, были созданы другие Big Data фреймворки, в частности, Apache Spark и [Flink](https://www.bigdataschool.ru/wiki/flink" \t "_blank).

Сплиты равняются размеру блока hdfs



Зависит от Resource manager



В частности, ответственность по управлению ресурсами кластера лежит на ResourceManager, а по планированию/координации жизненного цикла приложений – на ApplicationMaster. При этом каждый вычислительный узел разделен на произвольное количество контейнеров Container, содержащих предопределенное количество ресурсов: CPU, RAM и т.д., за которыми наблюдает менеджер узла (NodeManager)