Знакомство hadoop: MapReduce

MapReduce - парадигма параллельных вычислений и обработки больших наборов данных на компьютерных кластерах. Подробная информация о принципе работы описана тут (https://habr.com/ru/post/103467/).

Самая популярная реализация парадигмы находится в составе программного пакета Hadoop (http://hadoop.apache.org/). Существуют готовые сборки операционной системы с установленным Hadoop, дополненные программным обеспечением для управления вычислительным кластером и дополнительными сервисами. Например Cloudera, Hortonworks, MapR. Cloudera предоставляет стартовый набор mapreduce в виде образа виртуальной машины, которую студентам предлагается запустить. Пользователям Linux можно установить чистый hadoop и выполнить лабораторную без создания виртуальной машины. Последовательность действий будет примерно одинаковой.

Рассмотрим процесс работы на примере подсчета количества слов в текстовом файле:

1. Установите одно из средств виртуализации: VirtualBox (https://www.virtualbox.org/) / VMWare (https://www.vmware.com/ru/products/workstation-player.html) / KVM / Docker
2. Скачайте Cloudera QuickStart VM отсюда (https://www.cloudera.com/downloads/quickstart_vms.html) для своего средства виртуализации. Это образ заранее настроенной и сконфигурированной виртуальной машины.
3. В случае VirtualBox выберите `файл -> импорт конфигураций`. Затем в окне ввода выберите файл с виртуальной машиной Cloudera. После успешного импорта нажмите start. Дополнительная информация о виртуальных машинах доступна в официальной документации тут (https://docs.cloudera.com/documentation/enterprise/5-13-x/topics/quickstart_vm_administrative_information.html).
4. После запуска виртуальной машины откройте в ней терминал, и запустите cloudera-manager:

   ```
   sudo /home/cloudera/cloudera-manager --express --force
   ```

   В результате скрипт выдаст адрес панели управления, в которую можно перейти в браузере и ознакомиться с возможностями и сервисами `cloudera`.

5. Откройте текстовый редактор (*gedit*), и создайте файл `WordCount.java` в своей домашней директории с содержимым с сайта (https://archive.cloudera.com/cdh5/cdh/5/hadoop/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Source_Code).
6. Создайте папку build в домашней директории и скомпилируйте в неё java программу командой:

```
javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/*
WordCount.java -d build/ -Xlint
```

7. Создайте архив с программой java:

```
jar -cvf wordcount.jar -C build/ .
```

8. Создайте пару текстовых файлов для теста работы программы:

```
echo "Hello World Bye World" > file0
echo "Hello Hadoop Goodbye Hadoop" > file1
```

9. Теперь загрузите эти файлы в хранилище HDFS, предварительно создав папку:

```
hadoop fs -mkdir /user/cloudera/wordcount
hadoop fs -mkdir /user/cloudera/wordcount/input
hadoop fs -put file0 /user/cloudera/wordcount/input
hadoop fs -put file1 /user/cloudera/wordcount/input
```

10. Запустите программу на hadoop:

```
hadoop jar wordcount.jar  WordCount /user/cloudera/wordcount/input
/user/cloudera/wordcount/output
```

    В команде выше указывается название файла с программой, имя главного класса, адрес директории с данными на hdfs хранилище, и адрес выходной директории. Директория с результатом не должна существовать, иначе hadoop выдаст ошибку.
11. С помощью команд `hadoop fs` самостоятельно выведите на экран результат вычислений.
12. В комплекте с `hadoop` поставляется утилита `/usr/lib/hadoop-mapreduce/hadoop-streaming.jar` (документация) (https://hadoop.apache.org/docs/r1.2.1/streaming.html), которая перенаправляет входные данные шагов map и reduce в виде std-потоков в любую другую программу.

13. Напишите шаги map и reduce на языке Python и запустите задачу для подсчета слов. У вас должно получиться две программы. Одна (map) принимает текстовые строки из входного потока, разбивает их на сова и выдает результат в выходной поток. Вторая (reduce) принимает из входного потока сгруппированные по словам значения, считает их количество, и выдает в выходной поток.
14. Проверьте свою программу на наборе данных из этого файла (https://drive.google.com/open?id=1zO1pPPrqkPUEill474Kyo_UToZ71kJBb). Топ 5 самых популярных слов образуют цитату одного известного человека.