Урок 3. YARN & MR

Мог бы показать скринами но увы, ) в кластере нет данные файлы папки и т.д. так что, что есть то есть.

1. Запустить задачу из примеров, например, вычисление pi методом Монте-Карло

YARN\_EXAMPLES=/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop-mapreduce/  
yarn jar $YARN\_EXAMPLES/hadoop-mapreduce-examples.jar pi 16 1000000

Мы прейдем по ссылке: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop-mapreduce/

Дальше создаем переменную YARN\_EXAMPLES

По ссылке YARN\_EXAMPLES = /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/Hadoop-mapreduce/

Можно проверить через команду echo YARN\_EXAMPLES тогда нам должен указать путь.

Запускаем команду yarn jar $YARN\_EXAMPLES/hadoop-mapreduce-examples.jar pi 16 1000000

В результате должен создать 16 Map от #0 до #15

Проверим Ресурс менеджер должен появится наша задача

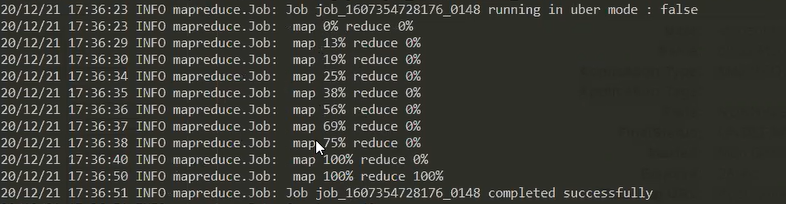


Рисунок – Процессы

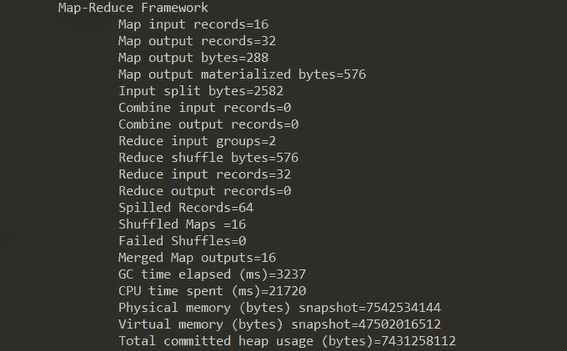


Рисунок – часть отчетов или логов



Рисунок – Число Пи 3,1415...

Создаем файл mapper форматом .py

Записываем

import sys

for line in sys.stdin:  
for word in line.split(" "):  
print(word + "\t" + str(1))

Создаем файл reducer форматом .py

import sys

prevkey = None  
cnt = 0

for line in sys.stdin:  
if(len(line.strip().split("\t")) > 1):  
key, value = line.strip().split("\t")  
#print(key + "\t" + value);  
if key != prevkey and prevkey is not None:  
print(prevkey + "\t" + str(cnt))  
cnt = 0  
cnt = cnt + int(value)  
prevkey = key

командой hdfs dfs -put сохраняем в hdfs

yarn jar /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop-mapreduce/Hadoop-streamin.jar -input /user/USERNAME/text\_example.txt -output result -file mapper.py -file reducer.py -mapper “python3 mapper.py” -reducer “python3 reducer.py”