Hadoop에 Spark 설치

개요

1. Spark를 사용하는 이유

1. 스파크는 HDFS에 저장된 데이터를 하둡 코어 라이브러리를 호출함으로써 메모리로 불러온 후, 변환 및 계산 등을 거쳐 최종 원하는 결과물을 산출한다.
2. Hadoop은 **디스크**로부터 map/reduce할 데이터를 불러오고, 처리 결과를 **디스크**로 쓴다. 따라서 데이터의 읽기/쓰기 속도는 느린 반면, 디스크 용량 만큼의 데이터를 한 번에 처리할 수 있다.
3. Spark는 **메모리**로부터 map/reduce할 데이터를 불러오고, 처리 결과를 **메모리**로 쓴다. 따라서 데이터의 읽기/쓰기 속도는 빠른 반면, 메모리 용량만큼의 데이터만 한 번의 처리할 수 있다. (메모리 용량보다 큰 데이터를 처리할 때는 과부하가 걸릴 수 있다)
4. 결론:
    - **메모리가 커버 가능**한 만큼의 데이터라면 **메모리 기반**이 유리 할 것이고, **메모리 용량 이상**의 데이터라면 **디스크 기반**이 유리하다.
    - plus, 기계학습이나 마이닝과 같은 반복 작업이 많을수록 메모리 기반이 유리하다.

2. Hadoop 위에 Spark를 설치해야 하는 이유

1. 스파크를 반드시 하둡과 함께 사용해야 할 필요는 없다.
2. 스파크는 하둡을 지원해서, 하둡이 사용하는 파일 시스템인 HDFS(Hadoop Distributed File System)의 데이터를 읽어올 수 있고 반대로 데이터를 쓸 수도 있다.
3. 따라서 매번 HDFS에 있는 파일을 Local로 옮기고 사용할 필요가 없어 데이터 처리 과정을 축소시킬 수 있다.

3. 참고

- https://wooono.tistory.com/50
- https://3months.tistory.com/511

Spark 설치 단계

1. hadoop 버전과 맞춰 spark 다운로드 주소 가져오기
2. ubuntu를 켜고 아래의 코드를 차례대로 입력

```
# $ cd Project/
$ wget https://dlcdn.apache.org/spark/spark-3.2.2/spark-3.2.2-bin-
hadoop3.2.tgz
$ tar zxf spark-3.2.2-bin-hadoop3.2.tgz
$ cp -r spark-3.2.2-bin-hadoop3.2 /home/hadoop/spark-3.2.2
$ cd
$ ll
```

3. `vi .bashrc`를 통해 .bashrc를 켜고, 맨 아래쪽에 있는 코드를 아래와 같이 작성한다.

```
# Automatically added
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME="/usr/local/hadoop"
export PATH="$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH"
export
PATH=$PATH:$HADOOP_HOME/sbin:~~~~:$FLUME_HOME/bin:$SPARK_HOME/bin:$SPARK_HOM
E/sbin
export SPARK_HOME=/home/hadoop/spark-3.2.2
```

4. 아래의 코드를 차례대로 입력하면 화면에 Spark 문구가 뜨고, 그러면 설치에 성공한 것이다.

```
$ source .bashrc
$ echo $SPARK_HOME
$ spark-submit --version
```

```
hadoop@hadoop:~$ vi .bashrc
hadoop@hadoop:~$ source .bashrc
hadoop@hadoop:~$ echo $SPARK_HOME
/home/hadoop/spark-3.2.2
hadoop@hadoop:~$ spark-submit --version
22/09/16 11:20:26 WARN Utils: Your hostname, hadoop resolves to a loopback addre
ss: 127.0.1.1; using 192.168.93.128 instead (on interface ens33)
22/09/16 11:20:26 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.2
      /_/

Using Scala version 2.12.15, OpenJDK 64-Bit Server VM, 1.8.0_342
Branch HEAD
Compiled by user centos on 2022-07-11T15:44:21Z
Revision 78a5825fe266c0884d2dd18cbca9625fa258d7f7
Url https://github.com/apache/spark
Type --help for more information.
```

5. spark-env.sh 안에 코드를 추가해야 하는데, 아래와 같이 입력한다.

```
$ cd spark-3.2.2/conf/
$ cp spark-env.sh.template spark-env.sh
$ vi spark-env.sh
```

```
# Options read by executors and drivers running inside the cluster
# - SPARK_LOCAL_IP, to set the IP address Spark binds to on this node
# - SPARK_PUBLIC_DNS, to set the public DNS name of the driver program
# - SPARK_LOCAL_DIRS, storage directories to use on this node for shuffle and RD
D data
# - MESOS_NATIVE_JAVA_LIBRARY, to point to your libmesos.so if you use Mesos
export SPARK_WORKER_INSTANCES=2
```

Spark 실행

1. 먼저 Hadoop을 실행시킨 다음
2. Spark를 실행시켜야 한다.

```
$ cd ~
$ start-dfs.sh
$ start-yarn.sh
$ jps   # 5개가 나와야 한다. (근데 나는 6개 나왔는데 그래도 괜찮았다.)
$ spark-shell   # spark 설치에 성공했을 때처럼 화면에 Spark 글씨가 크게 나온다.
```

Spark 예제

1. cd ~
2. 폴더를 새로 하나 만들고, word counting을 위한 txt 파일을 하나 만든다.
3. 아래 link에 있는 코드와 똑같이 연습해본다.

https://we-co.tistory.com/30

vmware에 python 설치하기

- 매번 위의 예제처럼 한 문장씩 입력할 수는 없기 때문에 실제 python 파일을 만들어서 돌려야 한다.
- vmware에서 새 폴더를 하나 만들고 예제 python 파일을 만들었는데, python3이 없다는 에러가 뜨면서 실행되지 않는다.
- cd ~을 해서 홈 디렉토리로 옮겨간 다음, 아래 코드를 입력해서 python을 설치하면 실행된다.
- (참고) python 파일을 실행할 때, 원래는 `python ex.py`로 해야 하지만, 여기서는 `python3 ex.py`로 해야 실행된다.

```
$ sudo apt-get update
$ sudo apt-get upgrade
$ sudo apt-get upgrade python3
$ pip3 --version    # command 'pip3' not found
$ sudo apt install python3-pip
$ sudo pip3 --version
$ python3
```

https://somjang.tistory.com/entry/PythonUbuntu%EC%97%90-Python-37-%EC%84%A4%EC%B9%98%ED%95%98%EA%B8%B0?category=345065

https://nyangnyangworld.tistory.com/3

PySpark

아래는 pyspark를 쓰는 예제이다.

코드를 입력하고 바로 실행하면 findspark, pyspark가 없다는 에러가 뜨는데, 그러면 맨 밑의 코드처럼 pip install 해주면 된다.

```
# input.py
import findspark
findspark.init()

import pyspark
from pyspark import SparkConf
from pyspark import SparkContext
from pyspark.sql import SQLContext
sc = pyspark.SparkContext()

# SparkContext version
print("spark Context version: ", sc.version)

# SparkContext python version
print("Spark Context Python version: ", sc.pythonVer)

# SparkContextMaster
print("Spark Context Master: ", sc.master)
```

```
$ python3 -m pip install findspark
$ python3 -m pip install pyspark

$ python3 input.py
```

https://0equal2.tistory.com/159

https://parkaparka.tistory.com/17