

Spark install & HBase + Phoenix connect

- 서버 6대를 통한 하둡 에코시스템 구축 가이드를 이어 진행한다는 가정
- spark master는 sn01, worker는 dn01, dn02, dn03
- hadoop 계정에서 운영
- 하둡, hbase, phoenix가 정상 작동되고 있음을 가정

Oracle VM VirtualBox 관리자

파일(F) 머신(M) 도움말(H)

도구

새로 만들기(N) 설정(S) 삭제 표시(H)

이름	상태	일반	시스템	미리 보기
an01 (hbase-phoenix-spark)	실행 중	이름: sn01 운영 체제: Red Hat (64-bit)	기본 메모리: 4096 MB 프로세서: 2 부팅 순서: 플로피, 광 디스크, 하드 디스크 가속: VT-X/AMD-V, 네스티드 페이징, PAE/NX, KVM 반가상화	
sn01 (hbase-phoenix-spark)	실행 중			
rm01 (hbase-phoenix-spark)	실행 중			
dn01 (hbase-phoenix-spark)	실행 중			
dn02 (hbase-phoenix-spark)	실행 중			
dn03 (hbase-phoenix-spark)	실행 중			

1. Spark install

- Spark 설치
- sn01 서버에서 설치, 설정 후 dn01, dn02, dn03서버에 배포

```
action server: sn01  
pwd: /home/hadoop
```

```
wget http://mirror.navercorp.com/apache/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz
```

```
tar xvfz spark-2.4.8-bin-hadoop2.7.tgz  
mv spark-2.4.8-bin-hadoop2.7 spark // 폴더명 변경
```

```
cd spark/conf // 경로 변경
```

1. Spark install

- **slaves 설정**

```
action server: sn01
pwd: /home/hadoop/spark/conf

vi slaves
dn01
dn02
dn03
```

- **spark-env.sh 설정**

```
action server: sn01
pwd: /home/hadoop/spark/conf

vi spark-env.sh
export JAVA_HOME=/opt/apps/jdk8
export SPARK_HOME=/home/hadoop/spark
export HADOOP_HOME=/home/hadoop/hadoop-3.1.2
export YARN_CONF_DIR=/home/hadoop/hadoop-3.1.2/etc/hadoop
export HADOOP_CONF_DIR=/home/hadoop/hadoop-3.1.2/etc/hadoop
```

1. Spark install

- **spark-defaults.conf 설정**

```
action server: sn01
pwd: /home/hadoop/spark/conf

vi spark-defaults.conf
spark.master yarn
spark.eventLog.enabled true
spark.eventLog.dir file:///home/hadoop/spark/eventLog
spark.history.fs.logDirectory file:///home/hadoop/spark/eventLog
cd .. // 경로 변경
```

- **이벤트로그 폴더 생성**

```
action server: sn01
pwd: /home/hadoop/spark

mkdir evenetLog
cd .. // 경로 변경
```

1. Spark install

- 재압축 후 배포

```
action server: sn01
```

```
pwd: /home/hadoop
```

```
tar cvfz spark.tar.gz spark
```

```
scp spark.tar.gz hadoop@dn01:/home/hadoop
```

```
scp spark.tar.gz hadoop@dn02:/home/hadoop
```

```
scp spark.tar.gz hadoop@dn03:/home/hadoop
```

- 압축 해제

```
action server: dn01, dn02, dn03
```

```
pwd: /home/hadoop
```

```
tar xvfz spark.tar.gz
```

1. Spark install

- 모든 하둡 클러스터의 yarn-site.xml 확인

```
action server: all
```

```
pwd: /home/hadoop/hadoop-3.1.2/etc/hadoop
```

```
<property>
  <name>yarn.nodemanager.pmem-check-enabled</name>
  <value>>false</value>
</property>
<property>
  <name>yarn.nodemanager.vmem-check-enabled</name>
  <value>>false</value>
</property>
```

위 부분이 없으면 하둡 종료, 파일 수정 후 다시 실행

1. Spark install

- **spark 환경변수 설정**

```
action server: sn01, dn01, dn02, dn03  
pwd: /root
```

```
vi /etc/profile.d/spark.sh
```

```
export SPARK_HOME=/home/hadoop/spark  
export PATH=$PATH:$SPARK_HOME/bin  
export PATH=$PATH:$SPARK_HOME/sbin
```

```
su - hadoop // 하둡 계정 재접속  
cd spark // 경로 변경
```

- **spark 실행**

```
action server: sn01  
pwd: /home/hadoop/spark
```

```
./sbin/start-master.sh  
./sbin/start-slaves.sh
```

```
jps // 상태 확인
```


- Master, Worker가 정상적으로 실행됨

sn01

```
[hadoop@sn01 spark]$ jps
7329 Master
1524 JournalNode
7399 Jps
1896 DFSZKFailoverController
1791 NameNode
[hadoop@sn01 spark]$
```

dn01

```
[hadoop@dn01 spark]$ jps
1921 DataNode
1429 NodeManager
3445 HRegionServer
4877 Jps
4782 Worker
[hadoop@dn01 spark]$
```

dn02

```
[hadoop@dn02 spark]$ jps
1952 DataNode
1426 NodeManager
3893 HRegionServer
5448 Worker
5546 Jps
[hadoop@dn02 spark]$
```


dn03

```
[hadoop@dn03 spark]$ jps
1425 NodeManager
3713 HRegionServer
5031 Worker
1917 DataNode
5133 Jps
[hadoop@dn03 spark]$
```

- 웹 UI 확인 (<http://192.168.56.101:8080>)

Spark Master at spark://sn01:7077

주의 요함 | 192.168.56.101:8080

 **Spark Master at spark://sn01:7077**

URL: spark://sn01:7077

Alive Workers: 3

Cores in use: 6 Total, 0 Used

Memory in use: 8.1 GB Total, 0.0 B Used

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20210615161538-192.168.56.103-43151	192.168.56.103:43151	ALIVE	2 (0 Used)	2.7 GB (0.0 B Used)
worker-20210615161538-192.168.56.104-35840	192.168.56.104:35840	ALIVE	2 (0 Used)	2.7 GB (0.0 B Used)
worker-20210615161538-192.168.56.105-34075	192.168.56.105:34075	ALIVE	2 (0 Used)	2.7 GB (0.0 B Used)

Running Applications (0)

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

2. Spark connect HBase & Phoenix

- phoenix jar 파일들을 모든 spark 클러스터 spark/jars 폴더에 복사

```
action server: an01  
pwd: /home/hadoop/phoenix
```

```
scp phoenix-5.0.0-HBase-2.0-client.jar hadoop@sn01:/home/hadoop/spark/jars  
scp phoenix-5.0.0-HBase-2.0-client.jar hadoop@dn01:/home/hadoop/spark/jars  
scp phoenix-5.0.0-HBase-2.0-client.jar hadoop@dn02:/home/hadoop/spark/jars  
scp phoenix-5.0.0-HBase-2.0-client.jar hadoop@dn03:/home/hadoop/spark/jars
```

```
scp phoenix-spark-5.0.0-HBase-2.0.jar hadoop@sn01:/home/hadoop/spark/jars  
scp phoenix-spark-5.0.0-HBase-2.0.jar hadoop@dn01:/home/hadoop/spark/jars  
scp phoenix-spark-5.0.0-HBase-2.0.jar hadoop@dn02:/home/hadoop/spark/jars  
scp phoenix-spark-5.0.0-HBase-2.0.jar hadoop@dn03:/home/hadoop/spark/jars
```

2. Spark connect HBase & Phoenix

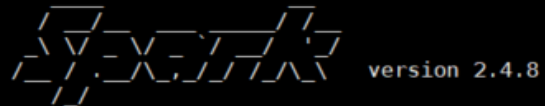
- pyspark 실행 및 phoenix & hbase 테이블 연동

```
action server: sn01  
pwd: /home/hadoop/spark
```

```
pyspark --master yarn // 파이스파크 실행
```

```
df = sqlContext.read ₩  
    .format("org.apache.phoenix.spark") ₩  
    .option("table", "SUBWAY2015") ₩  
    .option("zkUrl", "an01:2181") ₩  
    .load() // 데이터 프레임으로 로드
```

```
[hadoop@sn01 spark]$ pyspark --master yarn
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/spark/jars/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/spark/jars/phoenix-5.0.0-HBase-2.0-client.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2021-06-15 16:25:06,078 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2021-06-15 16:25:07,889 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
```



```
Using Python version 3.6.8 (default, Nov 16 2020 16:55:22)
SparkSession available as 'spark'.
```

```
>>> df = sqlContext.read \
...     .format("org.apache.phoenix.spark") \
...     .option("table", "SUBWAY2015") \
...     .option("zkUrl", "an01:2181") \
...     .load()
>>> df.show()
```

ROWKEY	use_dt	line_num	sub_sta_nm	ride_pasgr_num	alight_pasgr_num	work_dt
1	20150101	1호선	서울역	47071	40197	20151217
10	20150101	1호선	동묘앞	5345	5654	20151217
100	20150101	4호선	미아	8432	7836	20151217
1000	20150102	분당선	가천대	6443	7414	20151217
10000	20150119	7호선	가산디지털단지	53647	55118	20151217
100000	20150703	경부선	영등포	54720	59530	20151217
100001	20150703	경부선	신도림	4301	3888	20151217
100002	20150703	경부선	신길	12378	9771	20151217
100003	20150703	경부선	구로	24851	25737	20151217
100004	20150703	경부선	가산디지털단지	21109	23749	20151217
100005	20150703	경부선	금천구청	12441	11788	20151217
100006	20150703	경부선	석수	13112	11396	20151217
100007	20150703	경부선	관악	9645	8745	20151217
100008	20150703	경부선	안양	28733	28903	20151217
100009	20150703	경부선	영락	11290	11476	20151217
10001	20150119	7호선	철산	28304	28030	20151217
100010	20150703	경부선	금정	33140	32111	20151217
100011	20150703	경부선	군포	8302	7697	20151217
100012	20150703	경부선	의왕	9976	9296	20151217
100013	20150703	경부선	성균관대	17865	16407	20151217

only showing top 20 rows

```
>>> █
```