

2. Spark 설치 및 환경설정

목 차

1. Spark 소개
2. Spark 다운로드
3. Spark 설치
4. Spark 환경설정
5. Spark 실행
6. Spark SQL CLI 실행

1. Spark 소개

● 빅데이터 탐색에 활용하는 기술 – Spark 등장배경

1. 기존 RDBMS를 대신할 빅데이터 저장 매체 Hadoop 등장
2. Hadoop에서도 SQL을 사용하고자 만든 것이 바로 Hive
Hive를 통해 Hadoop에서도 SQL을 이용하여 DW 생성(편의성 제공)
Hive는 Hadoop의 MapReduce 방법을 이용하여 연산 수행

매 연산마다 다음과 같은 작업 반복

1. Disk에서 Memory로 연산에 필요한 Data 로딩
2. Memory에서 연산을 진행하고, 다시 Disk에 변경사항 저장

but 위 과정으로 불필요한 I/O 연산 많아지고, 처리 속도 떨어짐

3. hive 한계를 극복하기 위한 대안으로 Spark 등장
Spark는 한번에 연산을 수행할 Data를 모두 Memory에 불러온 후,
Memory에서 연산을 수행하기 때문에 Hive보다 훨씬 빠른 연산 가능

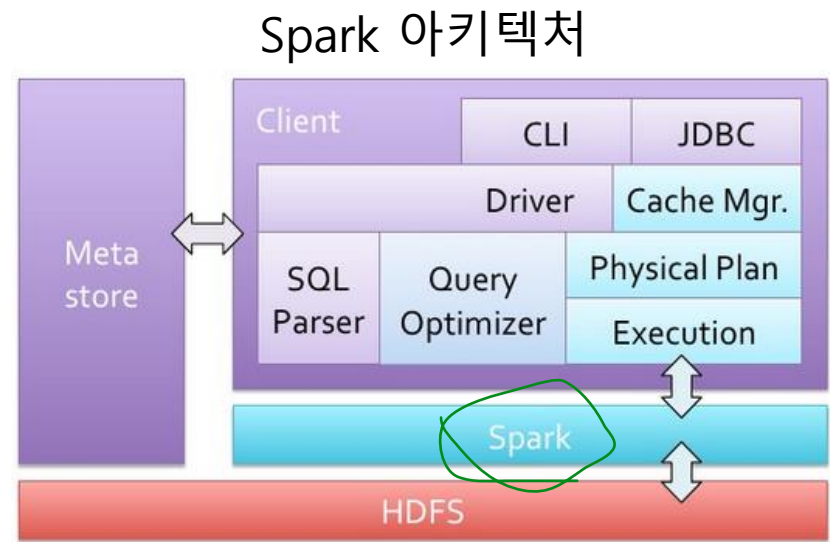
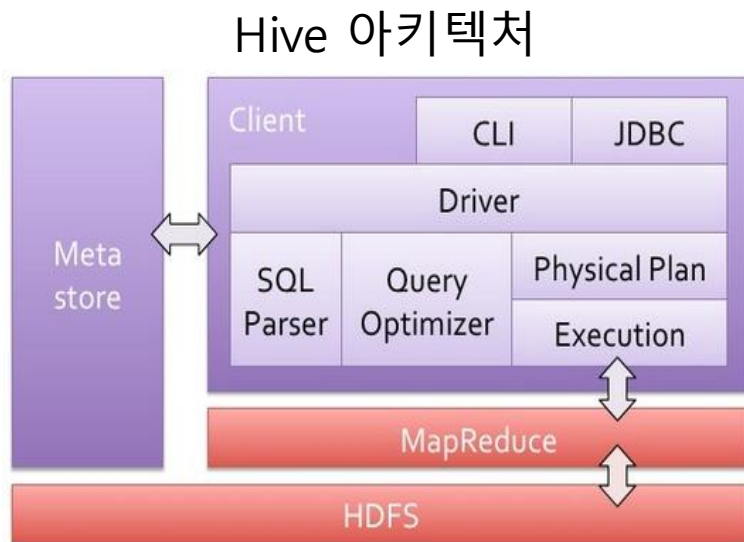
➤ Spark 소개

- 맵리듀스 코어를 그대로 사용하는 하이브는 성능면에서 여전히 느림
- 그로 인해 반복적인 대화형 연산 작업에서는 하이브가 적합하지 못함
- 이 단점을 극복한 고성능 인메모리 분석
- UC 버클리의 AMPLab에서 2009년 개발, 2010년 오픈 소스로 공개
- 2013년 6월 아파치 재단으로 이관되어 최상위 프로젝트
- 최근 빅데이터 분야에서 가장 핫한 기술 중 하나
- 데이터 가공 처리를 인메모리에서 수행함으로써 대용량 데이터 작업에도 빠른 성능을 보장

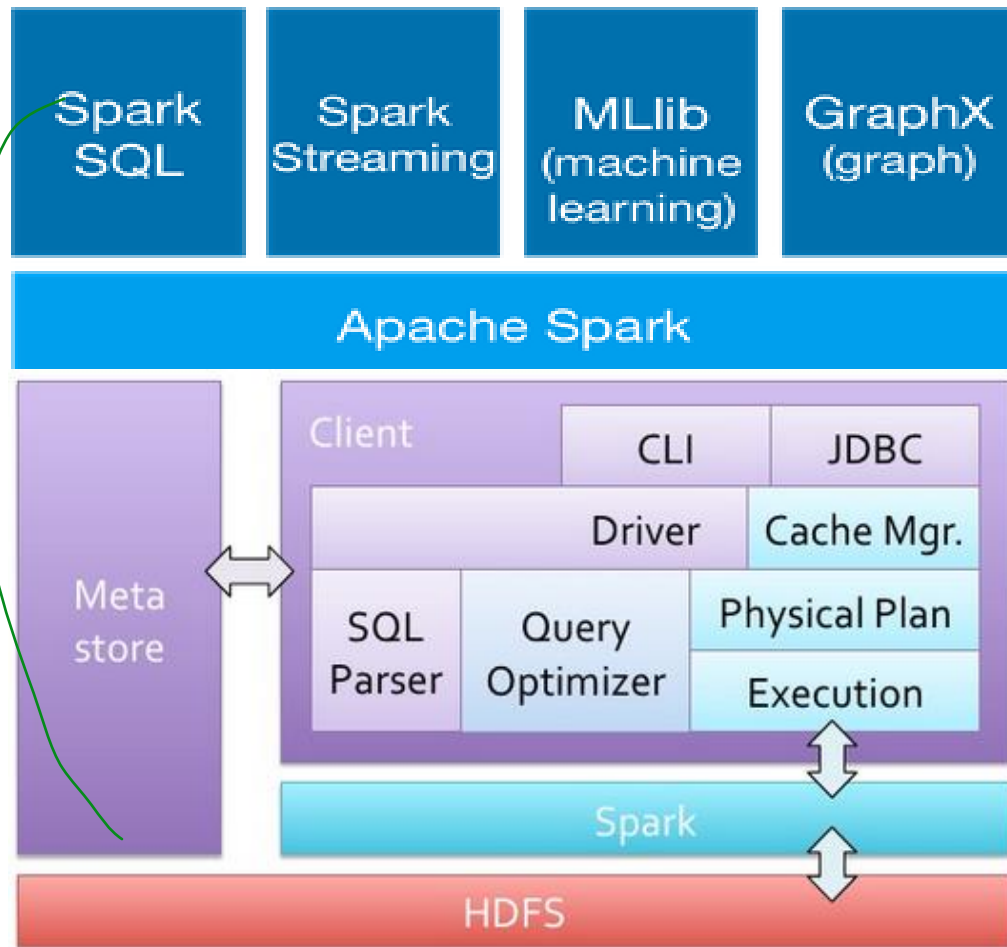
Hive vs Spark

➤ Hive vs Spark 아키텍처

- ✓ HiveQL은 MapReduce으로 변환하여 HDFS 데이터를 대상으로 DW를 생성하기 때문에 처리 속도가 느림



➤ Spark 아키텍처



- Spark SQL
Hive 대신 Spark SQL를 통해 MapReduce 없이 빠르게 처리
- Spark SQL CLI
HiveQL을 이용하여 테이블을 작성하거나 테이블에 데이터를 로드하고, 테이블에 대화식으로 쿼리를 발행하여 분산처리 구현
- Spark Streaming
스트림 데이터를 짧은 간격으로 읽어서 처리하는 처리하는 준 실시간 데이터 처리 방식
- MLib for machine learning
Classification, Regression, Clustering 등의 다양한 ML 알고리즘 지원
- GraphX
그래픽스 처리용 라이브러리 지원

2. Spark 설치

아래 사이트에서 Hadoop 버전에 맞는 Spark 버전을 찾아 설치를 진행한다.
Hadoop 3.3.6의 경우 Spark 3.4.4와 호환되기 때문에 3.4.4로 설치를 진행한다.

<http://spark.apache.org/downloads.html> 에서 다운로드 가능한 버전 확인

The screenshot shows the Apache Spark download page. The browser address bar displays `http://spark.apache.org/downloads.html`. The page header includes the Apache Spark logo and navigation links: Download, Libraries, Documentation, Examples, Community, Developers, and GitHub. The main content area is titled "Download Apache Spark™". It contains a list of steps for downloading Spark. Step 1, "Choose a Spark release:", has a dropdown menu showing "3.4.4 (Oct 27 2024)". A blue callout bubble with the number "1" and the text "버전 변경" (Change version) points to this dropdown. Step 2, "Choose a package type:", has a dropdown menu showing "Pre-built for Apache Hadoop 3.3 and later". Step 3, "Download Spark:", has the link "spark-3.4.4-bin-hadoop3.tgz" highlighted with a red box. A blue callout bubble with the number "2" and the text "링크 클릭으로 다운로드" (Download by clicking the link) points to this link. Step 4, "Verify this release using the 3.4.4 signatures, checksums, and release KEYS by following these procedures.", is partially visible. Below the steps, there is a note about Scala versions and a section titled "Link with Spark" which provides Maven coordinates for Spark artifacts.

1 버전 변경

Download Apache Spark™

1. Choose a Spark release: 3.4.4 (Oct 27 2024) ▾
2. Choose a package type: Pre-built for Apache Hadoop 3.3 and later ▾
3. Download Spark: [spark-3.4.4-bin-hadoop3.tgz](#)
4. Verify this release using the 3.4.4 signatures, checksums, and release KEYS by following these procedures.

Note that Spark 4 is pre-built with Scala 2.13, and Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

2 링크 클릭으로 다운로드

Link with Spark

Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.13
version: 4.0.0
```

1) Spark 다운로드 & 압축풀기(Master 작업)

The screenshot shows a terminal window with the title bar 'Home Linux Master Slave1'. The terminal prompt is 'hadoop@master:~'. A blue callout box 'Master에서 다운로드' points to the first command. The command is 'wget https://archive.apache.org/dist/spark/spark-3.4.4/spark-3.4.4-bin-hadoop3.tgz'. The output shows the file being downloaded from archive.apache.org. A second blue callout box '다운로드 파일 확인' points to the 'ls' command. The output of 'ls' shows the downloaded file 'spark-3.4.4-bin-hadoop3.tgz' among other files. A third blue callout box '압축풀기 : spark 설치' points to the 'tar -xvzf' command. The output of the command shows the contents of the Spark distribution being extracted.

```
[hadoop@master ~]$ wget https://archive.apache.org/dist/spark/spark-3.4.4/spark-3.4.4-bin-hadoop3.tgz
--2025-07-24 15:53:46-- https://archive.apache.org/dist/spark/spark-3.4.4/spark-3.4.4-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 388988563 (371M) [application/x-gzip]
Saving to: 'spark-3.4.4-bin-hadoop3.tgz'

100%===== 1.53MB/s in 4m 9s
2025-07-24 15:57:55 (1.49 MB/s) - 'spark-3.4.4-bin-hadoop3.tgz' saved [388988563/388988563]

[hadoop@master ~]$ ls
NASDAQ.zip          hadoop-3.3.6          jdk-8u461-linux-x64.tar.gz  문서          서식
apache-hive-3.1.3-bin  hadoop-3.3.6.tar.gz  spark-3.4.4-bin-hadoop3.tgz  바탕화면      음악
ap                  de
파일이(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@master ~]$ ls
NASDAQ.zip          hadoop-3.3.6          jdk-8u461-linux-x64.tar.gz  문서          서식
apache-hive-3.1.3-bin  hadoop-3.3.6.tar.gz  spark-3.4.4-bin-hadoop3.tgz  바탕화면      음악
ap                  de
파일이(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@master ~]$ tar -xvzf spark-3.4.4-bin-hadoop3.tgz
spark-3.4.4-bin-hadoop3/
spark-3.4.4-bin-hadoop3/ jars/
spark-3.4.4-bin-hadoop3/ jars/HikariCP-2.5.1.jar
spark-3.4.4-bin-hadoop3/ jars/JLargeArrays-1.5.jar
spark-3.4.4-bin-hadoop3/ jars/JTransforms-3.1.jar
spark-3.4.4-bin-hadoop3/ jars/RoaringBitmap-0.9.38.jar
spark-3.4.4-bin-hadoop3/ jars/ST4-4.0.4.jar
spark-3.4.4-bin-hadoop3/ jars/activation-1.1.1.jar
spark-3.4.4-bin-hadoop3/ jars/aircompressor-0.21.jar
```

● Soft link

```
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@master ~]$ ls
NASDAQ.zip                      hadoop-3.3.6                  jdk-8u461-linux-x64.tar.gz
apache-hive-3.1.3-bin            hadoop-3.3.6.tar.gz          spark-3.4.4-bin-hadoop3
apache-hive-3.1.3-bin.tar.gz     hadoopdata                   spark-3.4.4-bin-hadoop3.tgz
derby.log                       hive                          공개
[hadoop@master ~]$ ln -s spark-3.4.4-bin-hadoop3 spark
[hadoop@master ~]$
[hadoop@master ~]$ ls -l
합계 1598304
-rw-rw-r--. 1 hadoop hadoop 111104360 7월 24 15:17 NASDAQ.zip
drwxrwxr-x. 11 hadoop hadoop 4096 7월 23 18:18 apache-hive-3.1.3-bin
-rw-rw-r--. 1 hadoop hadoop 326940667 4월 9 2022 apache-hive-3.1.3-bin.tar.gz
-rw-rw-r--. 1 hadoop hadoop 22220 7월 24 14:48 derby.log
drwxr-xr-x. 11 hadoop hadoop 4096 7월 22 17:08 hadoop-3.3.6
-rw-rw-r--. 1 hadoop hadoop 730107476 6월 26 2023 hadoop-3.3.6.tar.gz
drwxrwxr-x. 3 hadoop hadoop 4096 7월 22 16:57 hadoopdata
lrwxrwxrwx. 1 hadoop hadoop 21 7월 23 17:53 hive -> apache-hive-3.1.3-bin
-rw-rw-r--. 1 hadoop hadoop 79436023 7월 21 17:50 jdk-8u461-linux-x64.tar.gz
lrwxrwxrwx. 1 hadoop hadoop 23 7월 24 16:08 spark -> spark-3.4.4-bin-hadoop3
drwxr-xr-x. 13 hadoop hadoop 4096 10월 21 2024 spark-3.4.4-bin-hadoop3
-rw-rw-r--. 1 hadoop hadoop 388988563 10월 21 2024 spark-3.4.4-bin-hadoop3.tgz
drwxr-xr-x. 2 hadoop hadoop 4096 7월 21 17:36 공개
```

소프트 링크

비디오

3. Spark 환경설정

1) .bash_profile 수정

```
hadoop@master:~  
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)  
[hadoop@master ~]$  
[hadoop@master ~]$  
[hadoop@master ~]$ vi .bash_profile
```

```
hadoop@master:~  
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)  
. ~/.bashrc  
fi  
  
# User specific environment and startup programs  
  
PATH=$PATH: $HOME/. local/bin: $HOME/bin  
  
export PATH  
  
export JAVA_HOME=/usr/local/jdk1.8.0_461  
export HADOOP_HOME=/home/hadoop/hadoop-3.3.6  
export PATH=$PATH: $JAVA_HOME/bin: $HADOOP_HOME/bin: $HADOOP_HOME/sbin  
  
# hive  
export HIVE_HOME=/home/hadoop/hive-3.12.0  
export PATH=$PATH: $HIVE_HOME/bin  
  
#spark  
export SPARK_HOME=/home/hadoop/spark  
export PATH=$PATH: $SPARK_HOME/bin  
  
: wq
```

SPARK_HOME 환경변수 추가 및 PATH 추가

.bash_profile 적용/테스트

```
hadoop@master:~/spark
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@master ~]$
[hadoop@master ~]$
[hadoop@master ~]$ vi .bash_profile
[hadoop@master ~]$ vi .bash_profile
[hadoop@master ~]$
[hadoop@master ~]$
[hadoop@master ~]$ source .bash_profile
[hadoop@master ~]$
[hadoop@master ~]$
[hadoop@master ~]$ cd $SPARK_HOME
[hadoop@master spark]$
[hadoop@master spark]$ pwd
/home/hadoop/spark
[hadoop@master spark]$
[hadoop@master spark]$
```

환경설정 파일 적용

Spark 홈 디렉터리 이동

Spark 홈 디렉터리 확인

2) Spark-env.sh 파일 생성/수정

hadoop@master:~/spark/conf

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

```
[hadoop@master ~]$  
[hadoop@master ~]$  
[hadoop@master ~]$  
[hadoop@master ~]$ cd $SPARK_  
[hadoop@master spark]$  
[hadoop@master spark]$ pwd  
/home/hadoop/spark  
[hadoop@master spark]$
```

Spark 환경설정 파일이 있는 디렉터리 이동

```
[hadoop@master spark]$ cd conf
```

```
[hadoop@master conf]$
```

```
[hadoop@master conf]$ ls
```

```
fairscheduler.xml.template metrics.properties.template  
log4j2.properties.template spark-defaults.conf.template
```

환경설정 파일 복사

```
[hadoop@master conf]$
```

```
[hadoop@master conf]$ cp spark-env.sh.template spark-env.sh
```

```
[hadoop@master conf]$
```

```
[hadoop@master conf]$ vi spark-env.sh
```

환경설정 파일 열기

3) Spark-env.sh 파일 생성/수정

```
hadoop@master:~/spark/conf
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
# - SPARK_HISTORY_OPTS, to set config properties only for the history server (e.g. "-Dx=y")
# - SPARK_SHUFFLE_OPTS, to set config properties only for the external shuffle service (e.g. "-Dx=y")
# - SPARK_DAEMON_JAVA_OPTS, to set config properties for all daemons (e.g. "-Dx=y")
# - SPARK_PUBLIC_DNS, to set the public dns name of the master or workers

# Generic options for the daemons used in the standalone deploy mode
# - SPARK_CONF_DIR      Alternate conf dir. (Default: ${SPARK_HOME}/conf)
# - SPARK_LOG_DIR       Where log files are stored. (Default: ${SPARK_HOME}/logs)
# - SPARK_PID_DIR       Where the pid file is stored. (Default: ${SPARK_HOME}/pid)
# - SPARK_IDENT_STRING  A string representing the daemon host.
# - SPARK_NICENESS       The scheduling priority for daemons. (Default: 0)
# - SPARK_NO_DAEMONIZE  Run the proposed command in the foreground. It will not output a PID file.

export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
```

Hadoop을 이용할 수 있도록 환경변수 추가

66, 0-1 바닥

파일 맨 밑에 밑의 3줄 추가
export HADOOP_HOME=/home/hadoop/hadoop-3.3.6
export HADOOP_CONF_DIR=\$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=\$HADOOP_HOME/lib/native

3) Spark-env.sh 파일 생성/수정

```
hadoop@master:~/spark/conf
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
# - SPARK_NICENESS      The scheduling priority for daemons. (Default: 0)
# - SPARK_NO_DAEMONIZE  Run the proposed command in the foreground. It will not output a PID file.
# Options for native BLAS, like Intel MKL, OpenBLAS, and so on.
# You might get better performance to enable these options if using native BLAS (see SPARK-21305).
# - MKL_NUM_THREADS=1   Disable multi-threading of Intel MKL
# - OPENBLAS_NUM_THREADS=1 Disable multi-threading of OpenBLAS

# Options for beeline
# - SPARK_BEELINE_OPTS, to set config properties only for the
# - SPARK_BEELINE_MEMORY, Memory for beeline (e.g. 1000M, 2G)

export HADOOP_HOME=/home/hadoop/hadoop-3.3.6
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native

:wq
```

Hadoop을 이용할 수 있도록
Hadoop 관련 환경변수 추가

5. Spark 실행

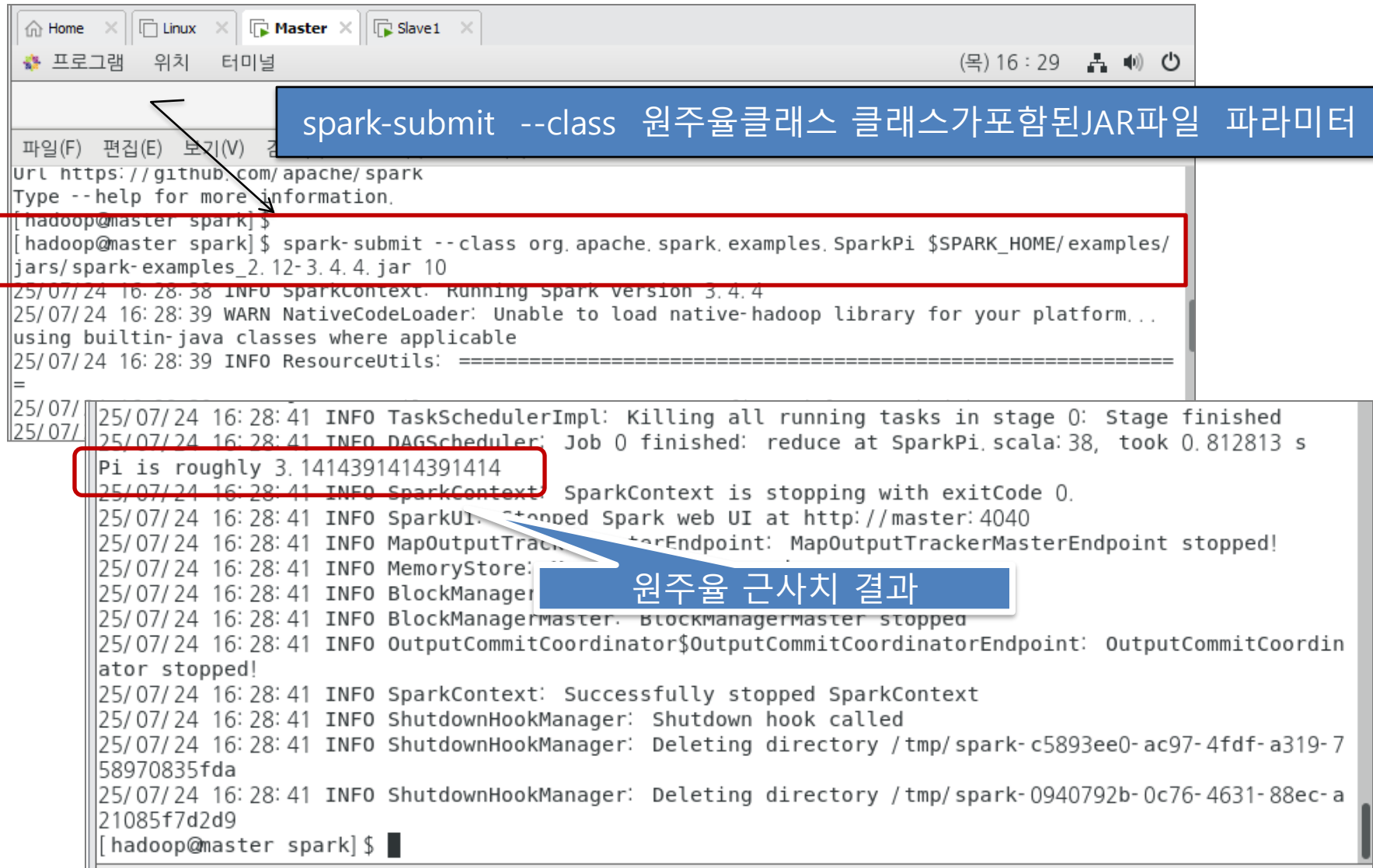
```
hadoop@master:~/spark

[hadoop@master conf]$
[hadoop@master conf]$
[hadoop@master conf]$ cd $SPARK_HOME
[hadoop@master spark]$
[hadoop@master spark]$
[hadoop@master spark]$ spark-submit --version
Welcome to
  ____
 /___\  _W W_  _W W_  _W W_  _W W_
/_ _/_/  _W W_  _W W_  _W W_  _W W_
/_ _/_/  _W W_  _W W_  _W W_  _W W_
/_ _/_/  _W W_  _W W_  _W W_  _W W_

version 3.4.4

Using Scala version 2.12.17, Java HotSpot(TM) 64-Bit Server VM, 1.8.0_461
Branch HEAD
Compiled by user ubuntu on 2024-10-21T02:09:45Z
Revision 6729992c76fc59ab07f63f97a9858691274447d0
Url https://github.com/apache/spark
Type --help for more information.
[hadoop@master spark]$
```

■ Spark 테스트 : 원주율 근사치 구하기



The screenshot shows a terminal window with tabs for 'Home', 'Linux', 'Master', and 'Slave1'. The terminal displays the execution of a Spark job to calculate pi. A blue callout box at the top points to the command line, explaining the parameters. A red box highlights the command and its immediate output. Another red box highlights the final result of the pi calculation. A final blue callout box points to the end of the log output.

```
spark-submit --class 원주율클래스 클래스가포함된JAR파일 파라미터

[hadoop@master spark]$ spark-submit --class org.apache.spark.examples.SparkPi $SPARK_HOME/examples/
jars/spark-examples_2.12-3.4.4.jar 10

25/07/24 16:28:38 INFO SparkContext: Running Spark version 3.4.4
25/07/24 16:28:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
25/07/24 16:28:39 INFO ResourceUtils: =====
=
25/07/24 16:28:41 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
25/07/24 16:28:41 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 0.812813 s
Pi is roughly 3.1414391414391414
25/07/24 16:28:41 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/07/24 16:28:41 INFO SparkUI: Stopped Spark web UI at http://master:4040
25/07/24 16:28:41 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/07/24 16:28:41 INFO MemoryStore: MemoryStore stopped
25/07/24 16:28:41 INFO BlockManager: BlockManager stopped
25/07/24 16:28:41 INFO BlockManagerMaster: BlockManagerMaster stopped
25/07/24 16:28:41 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordin
ator stopped!
25/07/24 16:28:41 INFO SparkContext: Successfully stopped SparkContext
25/07/24 16:28:41 INFO ShutdownHookManager: Shutdown hook called
25/07/24 16:28:41 INFO ShutdownHookManager: Deleting directory /tmp/spark-c5893ee0-ac97-4fdf-a319-7
58970835fda
25/07/24 16:28:41 INFO ShutdownHookManager: Deleting directory /tmp/spark-0940792b-0c76-4631-88ec-a
21085f7d2d9
[hadoop@master spark]$
```

6. Spark SQL CLI 실행

✓ Hive 서버 연동으로 Spark SQL 실행

1) Hive 서버 연동을 위해서 \$SPARK_HOME/conf 디렉터리에 Hive와 Hadoop 설정 파일 복사(hive-site.xml, core-site.xml, hdfs-site.xml)

```
hadoop@master:~/spark/conf
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
58970835fda
25/07/24 16:28:41 INFO ShutdownHook: directory /tmp/spark-0940792b-0c76-463
21085f7d2d9
[hadoop@master spark]$
[hadoop@master spark]$ cd conf/
[hadoop@master conf]$
[hadoop@master conf]$ pwd
/home/hadoop/spark/conf
[hadoop@master conf]$
[hadoop@master conf]$ cp $HIVE_HOME/conf/hive-site.xml .
[hadoop@master conf]$ cp $HADOOP_HOME/etc/hadoop/core-site.xml .
[hadoop@master conf]$ cp $HADOOP_HOME/etc/hadoop/hdfs-site.xml .
[hadoop@master conf]$ ls
core-site.xml          log4j2.properties.template  spark-env.sh.template
fairscheduler.xml.template  metrics.properties.template  workers.template
hdfs-site.xml          spark-defaults.conf.template
hive-site.xml          spark-env.sh
[hadoop@master conf]$
```

경로 이동

현재 위치에 Hadoop 설정파일 복사

hive 환경설정 파일 복사

hadoop 환경설정 파일 복사

2) Hadoop 실행 & spark-sql 실행

```
hadoop@master:~/spark
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
[hadoop@master conf]$
[hadoop@master conf]$
[hadoop@master conf]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/hadoop/hadoop-2.10.1/logs/hadoop-hadoop-namenode-master.out
slave1: datanode running as process 1797. Stop it first
Starting secondary namenodes [slave1]
slave1: secondary namenode running as process 1797. Stop it first
[hadoop@master conf]$
[hadoop@master conf]$ cd $SPARK_HOME
[hadoop@master spark]$
[hadoop@master spark]$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/07/24 16:41:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
25/07/24 16:41:02 WARN HiveConf: HiveConf of name hive.metastore.wm.default.pool.size does not exist
25/07/24 16:41:02 WARN HiveConf: HiveConf of name hive.llap.task.scheduler.preempt.independent does not exist
25/07/24 16:41:02 WARN HiveConf: HiveConf of name hive.llap.output.format.arrow does not exist
25/07/24 16:41:02 WARN HiveConf: HiveConf of name hive.tez.llap.min.reducer.per.executor does not exist
```

Hadoop 실행

Spark 홈 디렉터리 이동

Spark-sql 실행

Master - VMware Workstation 16 Player (Non-commercial use only)

Player | | | | |

프로그램 위치 터미널

파일(F) 편집(E) 보기(V) 검색(S) 터미널

24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.llap.task.scheduler.am.registry does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.druid.overlord.address.default does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.optimize.remove.sql.count.check does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.server2.webui.enable.cors does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.vectorized.row.serde.inputformat.excludes does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.query.reexecution.stats.cache.size does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.combine.equivalent.work.optimization does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.lock.query.string.max.length does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.llap.io.track.cache.usage does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.use.orc.codec.pool does not exist
24/07/09 15:49:36 WARN HiveConf: HiveConf of name hive.query.results.cache.max.size does not exist

Spark master: local[*] Application Id: local-1720507774183

spark-sql (default)>

hadoop@master:~

Spark-sql 프롬프트

Hive 설정 파일(hive-site.xml)에 정의된 설정 키가 현재 사용 중인 Hive 버전에서 지원되지 않거나 존재하지 않은 경우 나오는 경고 메시지 이므로 무시하고 넘어가도 됨

만약 경고 메시지를 차단하려면 다음 페이지를 참고한다.

- Hive 설정 파일(hive-site.xml)에서 경고 메시지 부분 태그 주석처리

Master - VMware Workstation 16 Player (Non-commercial use only)

Player | 프로그램 위치 터미널 ko (화) 16 : 03

hadoop@master:~/spark/conf

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

```
[hadoop@master conf]$  
[hadoop@master conf]$  
[hadoop@master conf]$ pwd  
/home/hadoop/spark/conf  
[hadoop@master conf]$  
[hadoop@master conf]$  
[hadoop@master conf]$ vi hive-site.xml
```

Spark의 conf 이동

Hive-site.xml 파일 열기

경고메시지 관련 태그 주석처리

```
</description>  
</property>  
  
<!--<property>  
  <name>hive.query.results.cache.max.size</name>  
  <value>2147483648</value>  
  <description>Maximum total size in bytes that the query results cache  
directory is allowed to use on the filesystem.</description>  
</property>-->  
  
<property>
```

6882, 0-1 99%

hadoop@master:~ hadoop@master:~/spark...

3) Spark SQL 실습 : iris 테이블 생성

sql에서는 ; 없으면 엔터하면 다음줄로 이동

hadoop@master:~/spark

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

테이블 생성

```
spark-sql (default)> create table iris_tab(  
    >         col1 float, col2 float, col3 float, col4 float, col5 string)  
    >         row format delimited fields terminated by ',' stored as textfile;
```

```
25/07/24 16:46:14 WARN SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.auth  
orization.manager is set to instance of HiveAuthorizerFactory.
```

```
25/07/24 16:46:14 WARN HiveConf: HiveConf of name hive.metastore.warehouse.dir does not exist
```

hadoop@master:~/spark

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

생성된 테이블 위치(HDFS)

```
25/07/24 16:46:14 WARN HiveConf: HiveConf of name hive.use.resource.coder.pool does not exist
```

```
25/07/24 16:46:14 WARN HiveConf: HiveConf of name hive.query.retry.times does not exist
```

```
25/07/24 16:46:14 WARN HiveConf: HiveConf of name hive.repl.bootstrap.parallelism does not  
exist
```

```
25/07/24 16:46:14 WARN HiveMetaStore: Location: hdfs://master:9000/user/hive/warehouse/iris_tab spec  
ified for non-external table: iris_tab
```

```
Time taken: 2.376 seconds
```

```
spark-sql (default)>
```

테이블 조회

```
> show tables;
```

```
25/07/24 16:49:12 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectExcept  
ion
```

```
iris_tab
```

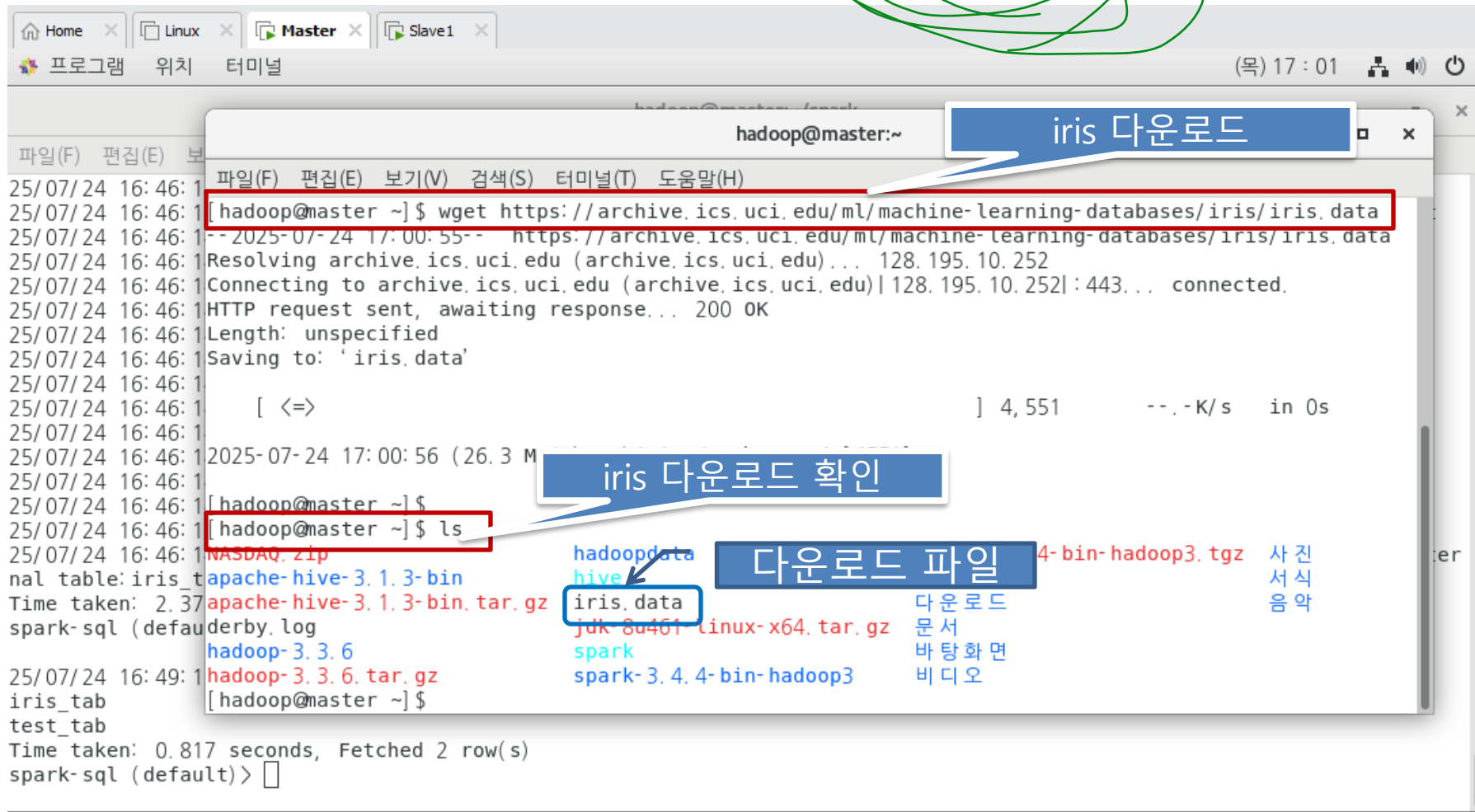
```
test_tab
```

```
Time taken: 0.817 seconds
```

```
spark-sql (default)>
```

테이블 조회 결과

● Spark SQL 실습 : iris 데이터셋 다운로드(새로운 터미널에서 작업)



```
hadoop@master:~$ wget https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
--2025-07-24 17:00:55-- https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
Resolving archive.ics.uci.edu (archive.ics.uci.edu)... 128.195.10.252
Connecting to archive.ics.uci.edu (archive.ics.uci.edu)|128.195.10.252|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified
Saving to: 'iris.data'

[ <=> ] 4,551 --.-K/s in 0s

2025-07-24 17:00:56 (26.3 M
hadoop@master:~$ ls
NASDAQ.zip      hadoopdata      4-bin-hadoop3.tgz  사진
apache-hive-3.1.3-bin  hive            다운로드          서식
apache-hive-3.1.3-bin.tar.gz  iris.data       문서              음악
derby.log       jdk-8u461-linux-x64.tar.gz  바탕화면
hadoop-3.3.6    spark           비디오
hadoop-3.3.6.tar.gz
hadoop@master:~$
```

● Spark SQL 실습 : iris-tab 테이블에 iris 데이터셋 삽입 & 조회

```
25/07/24 16:46:14 WARN HiveMetaStore: Location: hdfs://master:9000/user/hive/warehouse/iris_tab specif
nal table:iris_tab
Time taken: 2.376 seconds
spark-sql (default)>
> show tables;
25/07/24 16:49:12 WARN ObjectStore: Failed to get database glo
iris_tab
test_tab
Time taken: 0.817 seconds, Fetched 2 row(s)
spark-sql (default)> load data local inpath '/home/hadoop/iris.data' into table iris_tab;
Time taken: 1.663 seconds
spark-sql (default)>
```

데이터셋 삽입

hadoop@master:~/spark

파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)

```
spark-sql (default)> load data local inpath '/home/hadoop/iris.data' into table iris_tab;
Time taken: 1.663 seconds
spark-sql (default)>
> select * from iris_tab;
```

5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
4.3	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa

테이블 조회

spark
SQL
터미널에서

4) Hive metaStore에서 Spark 테이블 확인(새로운 터미널에서 작업)

The screenshot shows a terminal window titled 'hadoop@master:~'. The terminal displays the command `hdfs dfs -ls /user/hive/warehouse` and its output, which lists two files: `iris_tab` and `test_tab`. A red box highlights the command, and a blue callout box labeled 'HDFS에서 확인' points to it. Another blue callout box labeled '테이블 확인' points to the `iris_tab` entry in the output. The output shows the file permissions, owner, group, size, and timestamp for each file.

```
hadoop@master:~$ hdfs dfs -ls /user/hive/warehouse
Found 2 items
drwxrwxr-x - hadoop supergroup 0 2025-07-24 17:04 /user/hive/warehouse/iris_tab
drwxr-xr-x - hadoop supergroup 0 2025-07-24 14:50 /user/hive/warehouse/test_tab
```

Below the terminal window, a portion of a Spark SQL query result is visible, showing a table with columns for Iris setosa measurements.

setosa	sepal	petal	species
1.4	0.3	Iris-setosa	
1.7	0.3	Iris-setosa	
1.5	0.3	Iris-setosa	

● 테이블 삭제 및 종료

```
hadoop@master:~/spark
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
>
>
> drop table iris_tab;
Time taken: 0.719 seconds
spark-sql (default)>
> show tables;
test_tab
Time taken: 0.039 seconds, Fetched 1 row(s)
spark-sql (default)>
> quit
> ;
[hadoop@master spark]$
[hadoop@master spark]$
```

테이블 삭제

테이블 삭제 확인

Spark-sql 종료

이런 작업들도 가능

spark-sql> select * from iris_tab where col1 >= 6.5;

spark-sql> select col5, avg(col1), avg(col3) from iris_tab group by col5;