[1] SSH with pwd

linsong@linsongdeMacBook-Air ~ % **ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa**

Generating public/private rsa key pair.

Created directory '/Users/linsong/.ssh'.

Your identification has been saved in /Users/linsong/.ssh/id\_rsa.

Your public key has been saved in /Users/linsong/.ssh/id\_rsa.pub.

The key fingerprint is:

SHA256:cQzN6EvZMCSq8FU7Fal3AOoVQ76XN+3H4BwpMPiP7B0 linsong@linsongdeMacBook-Air.local

The key's randomart image is:

+---[RSA 3072]----+

| .Bo=\* |

| = B\*oo |

|. + \*o+\*o |

| o + ..+=\*o. . |

| o . .oSo+ = |

| o.+ \* + |

| o E + o |

| . . . . |

| . . |

+----[SHA256]-----+

linsong@linsongdeMacBook-Air ~ % **cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys**

linsong@linsongdeMacBook-Air ~ % **chmod 0600 ~/.ssh/authorized\_keys**

(base) linsong@linsongdeMacBook-Air /etc % **ssh localhost**

Last login: Tue Jan 26 09:48:59 2021

[2] Install Hadoop

linsong@linsongdeMacBook-Air ~ % **brew install hadoop**

linsong@linsongdeMacBook-Air ~ % **ls /usr/local/Cellar/hadoop**

3.3.0

[3] Java (Java SE Development Kit 8)

linsong@linsongdeMacBook-Air hadoop % **java -version**

java version "15.0.2" 2021-01-19

Java(TM) SE Runtime Environment (build 15.0.2+7-27)

Java HotSpot(TM) 64-Bit Server VM (build 15.0.2+7-27, mixed mode, sharing)

linsong@linsongdeMacBook-Air hadoop % **cd /Library/Java/JavaVirtualMachines**

linsong@linsongdeMacBook-Air JavaVirtualMachines % **ls -la**

total 0

drwxr-xr-x 6 root wheel 192 Jan 26 09:07 .

drwxr-xr-x 4 root wheel 128 Oct 24 2019 ..

drwxr-xr-x 3 root wheel 96 Apr 24 2020 adoptopenjdk-8.jdk

drwxr-xr-x 3 root wheel 96 Jan 25 21:15 jdk-15.0.2.jdk

drwxr-xr-x 3 root wheel 96 Jan 26 09:07 jdk1.8.0\_281.jdk

drwxr-xr-x@ 3 root wheel 96 Mar 5 2020 openjdk-14.0.1.jdk

linsong@linsongdeMacBook-Air JavaVirtualMachines % **/usr/libexec/java\_home -V**

Matching Java Virtual Machines (4):

15.0.2, x86\_64: "Java SE 15.0.2" /Library/Java/JavaVirtualMachines/jdk-15.0.2.jdk/Contents/Home

14.0.1, x86\_64: "OpenJDK 14.0.1" /Library/Java/JavaVirtualMachines/openjdk-14.0.1.jdk/Contents/Home

1.8.0\_281, x86\_64: "Java SE 8" /Library/Java/JavaVirtualMachines/jdk1.8.0\_281.jdk/Contents/Home

1.8.0\_252, x86\_64: "AdoptOpenJDK 8" /Library/Java/JavaVirtualMachines/adoptopenjdk-8.jdk/Contents/Home

/Library/Java/JavaVirtualMachines/jdk-15.0.2.jdk/Contents/Home

Open **~/.bash\_profile**, and add:

export JAVA\_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0\_281.jdk/Contents/Home

export PATH=$JAVA\_HOME/bin:$PATH

linsong@linsongdeMacBook-Air JavaVirtualMachines % **echo $JAVA\_HOME**

linsong@linsongdeMacBook-Air JavaVirtualMachines % **source ~/.bash\_profile**

(base) linsong@linsongdeMacBook-Air JavaVirtualMachines % **echo $JAVA\_HOME**

/Library/Java/JavaVirtualMachines/jdk1.8.0\_281.jdk/Contents/Home

[4] Configuration

linsong@linsongdeMacBook-Air ~ % **cd /usr/local/Cellar/hadoop/3.3.0/libexec/etc/Hadoop**

(base) linsong@linsongdeMacBook-Air hadoop % **/usr/libexec/java\_home**

/Library/Java/JavaVirtualMachines/jdk-15.0.2.jdk/Contents/Home

**hadoop-env.sh**

export JAVA\_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0\_281.jdk/Contents/Home

**core-site.xml**

<configuration>

<!--namenode address-->

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

**hdfs-site.xml**

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9000</value>

</property>

<property>

<name>dfs.http.address</name>

<value>localhost:9870</value>

</property>

<property>

<name>dfs.permissions</name>

<value>false</value> <!--turn off firewall-->

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>/Users/linsong/Documents/2021/BigData/hadoop/dfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>/Users/linsong/Documents/2021/BigData/hadoop/dfs/data1/hadoop</value>

</property>

</configuration>

**mapred-site.xml**

<configuration>

<!--hadoop MR(Map/Reduce) over YAR-->

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

**yarn-site.xml**

<configuration>

<!--nomenodeManager shuffle-->

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<!--NodeManagers env variables-->

<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

</configuration>

[5] Format

linsong@linsongdeMacBook-Air ~ % **cd /usr/local/Cellar/hadoop/3.3.0/**

**Mac: 选择系统偏好设置->选择共享->点击远程登录**

linsong@linsongdeMacBook-Air 3.3.0 % **bin/hdfs namenode -format**

linsong@linsongdeMacBook-Air 3.3.0 % **sbin/start-yarn.sh**

Starting resourcemanager

resourcemanager is running as process 38270. Stop it first.

Starting nodemanagers

localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.

Open: <http://localhost:8088/cluster>

[6] Install Hive

linsong@linsongdeMacBook-Air 3.3.0 % **brew install hive**

linsong@linsongdeMacBook-Air 3.3.0 % **cd /usr/local/Cellar/hadoop/3.3.0/**

linsong@linsongdeMacBook-Air 3.3.0 % **sbin/start-all.sh**

linsong@linsongdeMacBook-Air 3.3.0 % **jps**

40160 SecondaryNameNode

38611 NodeManager

40029 DataNode

40510 Jps

38270 ResourceManager

[7] MySql

Root: 12345678

Open **~/.bash\_profile**, and add:

export PATH=${PATH}:/usr/local/mysql/bin

linsong@linsongdeMacBook-Air JavaVirtualMachines % **source ~/.bash\_profile**

(base) linsong@linsongdeMacBook-Air bin % **mysql -u root -p**

Enter password:

Welcome to the MySQL monitor. Commands end with ; or \g.

mysql> **create database metastore;**

Query OK, 1 row affected (0.26 sec)

Add **hive\_site.xml** to /usr/local/Cellar/hive/3.1.2\_2/libexec/conf:

|  |
| --- |
| <configuration>  　　<property>  <name>hive.metastore.local</name>  <value>true</value>  </property>  <property>  <name>javax.jdo.option.ConnectionURL</name>  <value>jdbc:mysql://localhost/metastore</value>  </property>  <property>  <name>javax.jdo.option.ConnectionDriverName</name>  <value>com.mysql.jdbc.Driver</value>  </property>  <property>  <name>javax.jdo.option.ConnectionUserName</name>  <value>root</value>  </property>  　　<property>  <name>javax.jdo.option.ConnectionPassword</name>  <value>12345678</value>  </property>  <property>  <name>hive.exec.local.scratchdir</name>  <value>/tmp/hive</value>  </property>  <property>  <name>hive.querylog.location</name>  <value>/tmp/hive</value>  </property>  <property>  <name>hive.downloaded.resources.dir</name>  <value>/tmp/hive</value>  </property>  <property>  <name>hive.server2.logging.operation.log.location</name>  <value>/tmp/hive</value>  </property>  </configuration> |

Download Connector/J 8.0.23 with platform independent, place **mysql-connector-java-8.0.23.jar** under **/usr/local/Cellar/hive/3.1.2\_2/libexec/lib**

linsong@linsongdeMacBook-Air 3.3.0 % **schematool -initSchema -dbType mysql**

mysql> show databases;

mysql> show tables;

[8] Verification

(base) linsong@linsongdeMacBook-Air local % **hadoop version**

Hadoop 3.3.0

Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r aa96f1871bfd858f9bac59cf2a81ec470da649af

Compiled by brahma on 2020-07-06T18:44Z

Compiled with protoc 3.7.1

From source with checksum 5dc29b802d6ccd77b262ef9d04d19c4

This command was run using /usr/local/Cellar/hadoop/3.3.0/libexec/share/hadoop/common/hadoop-common-3.3.0.jar

<http://localhost:9870/>

<http://localhost:8088/cluster>

(base) linsong@linsongdeMacBook-Air 3.3.0 % **hive**

hive> **show databases;**

OK

default

Time taken: 6.103 seconds, Fetched: 1 row(s)

hive> **show tables;**

OK

Time taken: 0.438 seconds

linsong@linsongdeMacBook-Air 3.3.0 % **hadoop fs -ls /**

2021-01-26 12:00:19,026 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 2 items

drwx-wx-wx - linsong supergroup 0 2021-01-26 11:47 /tmp

drwxr-xr-x - linsong supergroup 0 2021-01-26 11:47 /user

linsong@linsongdeMacBook-Air **bin % hadoop jar /usr/local/Cellar/hadoop/3.3.0/libexec/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar pi 10 10**

Job Finished in 92.717 seconds

Estimated value of Pi is 3.20000000000000000000

[9] Spark

Add below to ~/.bash\_profile:

CLASSPATH=$JAVA\_HOME/lib/tools.jar:$JAVA\_HOME/lib/dt.jar:.

export CLASSPATH

(base) linsong@linsongdeMacBook-Air ~ % **source ~/.bash\_profile**

(base) linsong@linsongdeMacBook-Air ~ % **java -version**

java version "1.8.0\_281"

Java(TM) SE Runtime Environment (build 1.8.0\_281-b09)

Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)

(base) linsong@linsongdeMacBook-Air ~ % **python3**

Python 3.8.5 (default, Sep 4 2020, 02:22:02)

[Clang 10.0.0 ] :: Anaconda, Inc. on darwin

Type "help", "copyright", "credits" or "license" for more information.

>>>

(base) linsong@linsongdeMacBook-Air ~ % **Scala -version**

Scala code runner version 2.13.2 -- Copyright 2002-2020, LAMP/EPFL and Lightbend, Inc.

Add below to ~/.bash\_profile:

export SCALA\_HOME=/usr/local/Cellar/scala/2.13.2

export PATH=$PATH:$SCALA\_HOME/bin

(base) linsong@linsongdeMacBook-Air 2.13.2 % **source ~/.bash\_profile**

linsong@linsongdeMacBook-Air ~ % **scala**

Welcome to Scala 2.13.2 (OpenJDK 64-Bit Server VM, Java 15.0.1).

Type in expressions for evaluation. Or try :help.

scala> **util.Properties.versionString**

val **res0**: **String** = version 2.13.2

scala> **:q**

linsong@linsongdeMacBook-Air ~ %

Download spark-3.0.1-bin-hadoop3.2.tgz to /usr/local and rename to spark3.0.1, and add below to

export SPARK\_HOME=/usr/local/spark3.0.1

export PATH=$PATH:$SPARK\_HOME/bin

(base) linsong@linsongdeMacBook-Air 2.13.2 % **source ~/.bash\_profile**

(base) linsong@linsongdeMacBook-Air Cellar % **cd /usr/local/spark3.0.1/conf/**

(base) linsong@linsongdeMacBook-Air conf % **cp spark-env.sh.template spark-env.sh**

Add below to **spark-env.sh**:

export SCALA\_HOME=/usr/local/Cellar/scala

export SPARK\_MASTER\_IP=localhost

export SPARK\_WORKER\_MEMORY=4g

(base) linsong@linsongdeMacBook-Air spark3.0.1 % **spark-shell**

2021-01-26 20:45:29,540 WARN util.Utils: Your hostname, linsongdeMacBook-Air.local resolves to a loopback address: 127.0.0.1; using 192.168.18.3 instead (on interface en0)

2021-01-26 20:45:29,552 WARN util.Utils: Set SPARK\_LOCAL\_IP if you need to bind to another address

2021-01-26 20:45:32,403 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).

Spark context Web UI available at http://192.168.18.3:4040

Spark context available as 'sc' (master = local[\*], app id = local-1611665154817).

Spark session available as 'spark'.

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\ version 3.0.1

/\_/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0\_281)

Type in expressions to have them evaluated.

Type :help for more information.

scala> import org.apache.spark.{SparkConf, SparkContext}

import org.apache.spark.{SparkConf, SparkContext}

scala> SparkContext

**res0**: **org.apache.spark.SparkContext.type** = org.apache.spark.SparkContext$@1780bc15

linsong@linsongdeMacBook-Air spark3.0.1 % **pwd**

/usr/local/spark3.0.1

linsong@linsongdeMacBook-Air spark3.0.1 % **sbin/start-all.sh**

starting org.apache.spark.deploy.master.Master, logging to /usr/local/spark3.0.1/logs/spark-linsong-org.apache.spark.deploy.master.Master-1-linsongdeMacBook-Air.local.out

localhost: starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark3.0.1/logs/spark-linsong-org.apache.spark.deploy.worker.Worker-1-linsongdeMacBook-Air.local.out

And then:

(base) linsong@linsongdeMacBook-Air spark3.0.1 % **spark-shell**

scala> **val textFile = sc.textFile("file:///usr/local/spark3.0.1/README.md")**

**textFile**: **org.apache.spark.rdd.RDD[String]** = file:///usr/local/spark3.0.1/README.md MapPartitionsRDD[1] at textFile at <console>:24

scala> **textFile.count()**

**res0**: **Long** = 108

scala> **textFile.first()**

**res1**: **String** = # Apache Spark

scala> **val linesWithSpark = textFile.filter(line => line.contains("Spark"))**

**linesWithSpark**: **org.apache.spark.rdd.RDD[String]** = MapPartitionsRDD[2] at filter at <console>:25

scala> **linesWithSpark.count()**

**res2**: **Long** = 19

scala> **textFile.map(line => line.split(" ").size).reduce((a, b) => if(a>b) a else b)**

**res3**: **Int** = 16

scala> **import java.lang.Math**

import java.lang.Math

scala> **textFile.map(line => line.split(" ").size).reduce((a, b) => Math.max(a, b))**

**res4**: **Int** = 16

scala> **val wordCounts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey((a,b) => a+ b)**

**wordCounts**: **org.apache.spark.rdd.RDD[(String, Int)]** = ShuffledRDD[7] at reduceByKey at <console>:26

scala> **wordCounts.collect()**

**res5**: **Array[(String, Int)]** = Array((package,1), (this,1), (integration,1), (Python,2), (cluster.,1), (its,1), ([run,1), (There,1), (general,2), (have,1), (pre-built,1), (Because,1), (YARN,,1), (locally,2), (changed,1), (locally.,1), (several,1), (only,1), (Configuration,1), (This,2), (basic,1), (first,1), (learning,,1), (documentation,3), (graph,1), (Hive,2), (info,1), (["Specifying,1), ("yarn",1), ([params]`.,1), ([project,1), (prefer,1), (SparkPi,2), (engine,2), (version,1), (file,1), (documentation,,1), (MASTER,1), (example,3), (are,1), (systems.,1), (params,1), (scala>,1), (DataFrames,,1), (provides,1), (refer,2), (configure,1), (Interactive,2), (R,,1), (can,6), (build,3), (when,1), (easiest,1), (Maven](https://maven.apache.org/).,1), (Apache,1), (guide](ht...

[10] Spark from Hive

linsong@linsongdeMacBook-Air spark3.0.1 % **bin/run-example SparkPi 2>&1 | grep "Pi is"**

Pi is roughly 3.1359356796783984

linsong@linsongdeMacBook-Air conf % **pwd**

/usr/local/spark3.0.1/conf

linsong@linsongdeMacBook-Air conf % **cp /usr/local/Cellar/hive/3.1.2\_2/libexec/conf/hive-site.xml .**

Below looks good, supporting hive from local:

scala> **import org.apache.spark.sql.hive.HiveContext**

import org.apache.spark.sql.hive.HiveContext

--1-- Create database in Hive

hive> **create database if not exists sparktest;**

OK

Time taken: 0.595 seconds

hive> **show databases;**

OK

default

sparktest

Time taken: 0.402 seconds, Fetched: 2 row(s)

hive> **use sparktest;**

OK

hive> **create table if not exists sparktest.student(id int,name string,gender string,age int);**

hive> **insert into student values(1,'Song','M',39);**

Query ID = linsong\_20210126222200\_41b9e97e-d514-467d-9e95-9dbb3e20445a

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks determined at compile time: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1611634845458\_0002, Tracking URL = http://localhost:8088/proxy/application\_1611634845458\_0002/

Kill Command = /usr/local/Cellar/hadoop/3.3.0/libexec/bin/mapred job -kill job\_1611634845458\_0002

**[Problem solving]**

When applying insertion from above, getting hang from Spark, here is the solution

(base) linsong@linsongdeMacBook-Air spark3.0.1 % **hadoop dfsadmin -report**

Configured Capacity: 0 (0 B)

Present Capacity: 0 (0 B)

DFS Remaining: 0 (0 B)

DFS Used: 0 (0 B)

DFS Used%: 0.00%

Replicated Blocks:

Under replicated blocks: 0

Blocks with corrupt replicas: 0

Missing blocks: 0

Missing blocks (with replication factor 1): 0

Low redundancy blocks with highest priority to recover: 0

Pending deletion blocks: 0

Erasure Coded Block Groups:

Low redundancy block groups: 0

Block groups with corrupt internal blocks: 0

Missing block groups: 0

Low redundancy blocks with highest priority to recover: 0

Pending deletion blocks: 0

From above, this is caused by inconsistent spaceID from previous -format to the namenode multiple times; to solve it:

<1> $./stop-all.sh

<2> remove all contents for below two directories under hdfs-site.xml

<property>

<name>dfs.namenode.name.dir</name>

<value>/Users/linsong/Documents/2021/BigData/hadoop/dfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>/Users/linsong/Documents/2021/BigData/hadoop/dfs/data1/hadoop</value>

</property>

<3> ./hadoop namenode -format (optional, no action on this one to solve the issue)

<4> $./start-all.sh

hive> insert into student values(2,'Jay','F',29);

Query ID = linsong\_20210127090234\_77729028-bff6-498e-a172-7043ac825e89

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks determined at compile time: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1611709278755\_0001, Tracking URL = http://linsongdeMacBook-Air.local:8088/proxy/application\_1611709278755\_0001/

Kill Command = /usr/local/Cellar/hadoop/3.3.0/libexec/bin/mapred job -kill job\_1611709278755\_0001

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2021-01-27 09:03:17,859 Stage-1 map = 0%, reduce = 0%

2021-01-27 09:03:34,802 Stage-1 map = 100%, reduce = 0%

2021-01-27 09:03:46,738 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_1611709278755\_0001

Stage-4 is selected by condition resolver.

Stage-3 is filtered out by condition resolver.

Stage-5 is filtered out by condition resolver.

Moving data to directory hdfs://localhost:9000/user/hive/warehouse/sparktest.db/student/.hive-staging\_hive\_2021-01-27\_09-02-34\_445\_1957612793997729063-1/-ext-10000

Loading data to table sparktest.student

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 HDFS Read: 17707 HDFS Write: 314 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

Time taken: 80.536 seconds

hive> select \* from student;

OK

1 Song M 39

2 Jay F 29

Time taken: 1.169 seconds, Fetched: 2 row(s)

[**Problem solving**]

/usr/local/spark3.0.1/conf/hive-site.xml

Comment off below:

<!-- 　　<property>

<name>hive.metastore.local</name>

<value>true</value>

</property> -->

/usr/local/Cellar/hive/3.1.2\_2/libexec/conf/hive-site.xml

Comment off below:

<!-- 　　<property>

<name>hive.metastore.local</name>

<value>true</value>

</property> -->

linsong@linsongdeMacBook-Air 3.3.0 % **cp -r /usr/local/Cellar/hive/3.1.2\_2/libexec/lib/mysql-connector-java-8.0.23.jar /usr/local/spark3.0.1/jars**

(base) linsong@linsongdeMacBook-Air spark3.0.1 % **spark-shell**

2021-01-27 20:28:28,850 WARN util.Utils: Your hostname, linsongdeMacBook-Air.local resolves to a loopback address: 127.0.0.1; using 192.168.18.3 instead (on interface en0)

2021-01-27 20:28:28,852 WARN util.Utils: Set SPARK\_LOCAL\_IP if you need to bind to another address

2021-01-27 20:28:30,962 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).

2021-01-27 20:28:54,455 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

Spark context Web UI available at http://192.168.18.3:4041

Spark context available as 'sc' (master = local[\*], app id = local-1611750534826).

Spark session available as 'spark'.

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\ version 3.0.1

/\_/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0\_281)

Type in expressions to have them evaluated.

Type :help for more information.

scala> **import org.apache.spark.sql.hive.HiveContext**

import org.apache.spark.sql.hive.HiveContext

scala> **val hiveCtx = new HiveContext(sc)**

warning: there was one deprecation warning (since 2.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'

**hiveCtx**: **org.apache.spark.sql.hive.HiveContext** = org.apache.spark.sql.hive.HiveContext@1cea4fdf

scala> **val studentRDD = hiveCtx.sql("select \* from sparktest.student").rdd**

2021-01-27 20:30:09,559 WARN conf.HiveConf: HiveConf of name hive.stats.jdbc.timeout does not exist

2021-01-27 20:30:09,561 WARN conf.HiveConf: HiveConf of name hive.stats.retries.wait does not exist

Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.

2021-01-27 20:30:17,610 WARN metastore.ObjectStore: Failed to get database global\_temp, returning NoSuchObjectException

**studentRDD**: **org.apache.spark.rdd.RDD[org.apache.spark.sql.Row]** = MapPartitionsRDD[6] at rdd at <console>:26

scala> **studentRDD.foreach(t => println("Name:"+t(1)+",Gender:"+t(2)+",Age:"+t(3)))**

Name:Jay,Gender:F,Age:29 (0 + 2) / 2]

Name:Song,Gender:M,Age:39

scala> **import org.apache.spark.sql.{SQLContext, Row}**

import org.apache.spark.sql.{SQLContext, Row}

scala> **import org.apache.spark.sql.types.{StringType, IntegerType, StructField, StructType}**

import org.apache.spark.sql.types.{StringType, IntegerType, StructField, StructType}

scala> **import org.apache.spark.sql.hive.HiveContext**

import org.apache.spark.sql.hive.HiveContext

scala> **val hiveCtx = new HiveContext(sc)**

warning: there was one deprecation warning (since 2.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'

**hiveCtx**: **org.apache.spark.sql.hive.HiveContext** = org.apache.spark.sql.hive.HiveContext@5c1b20e9

scala> **val studentRDD = sc.parallelize(Array("3 Roger M 60","4 Tim M 27")).map(\_.split(" "))**

**studentRDD**: **org.apache.spark.rdd.RDD[Array[String]]** = MapPartitionsRDD[8] at map at <console>:29

scala> **val schema = StructType(List(StructField("id", IntegerType, true),StructField("name", StringType, true),StructField("gender", StringType, true),StructField("age", IntegerType, true)))**

**schema**: **org.apache.spark.sql.types.StructType** = StructType(StructField(id,IntegerType,true), StructField(name,StringType,true), StructField(gender,StringType,true), StructField(age,IntegerType,true))

scala> **val rowRDD = studentRDD.map(p => Row(p(0).toInt, p(1).trim, p(2).trim, p(3).toInt))**

**rowRDD**: **org.apache.spark.rdd.RDD[org.apache.spark.sql.Row]** = MapPartitionsRDD[9] at map at <console>:30

scala> **val studentDataFrame = hiveCtx.createDataFrame(rowRDD, schema)**

**studentDataFrame**: **org.apache.spark.sql.DataFrame** = [id: int, name: string ... 2 more fields]

scala> **studentDataFrame.registerTempTable("tempTable")**

warning: there was one deprecation warning (since 2.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'

scala> **hiveCtx.sql("insert into sparktest.student select \* from tempTable")**

2021-01-27 21:01:06,402 WARN session.SessionState: METASTORE\_FILTER\_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.

2021-01-27 21:01:06,555 WARN conf.HiveConf: HiveConf of name hive.internal.ss.authz.settings.applied.marker does not exist

2021-01-27 21:01:06,555 WARN conf.HiveConf: HiveConf of name hive.stats.jdbc.timeout does not exist

2021-01-27 21:01:06,557 WARN conf.HiveConf: HiveConf of name hive.stats.retries.wait does not exist

**res2**: **org.apache.spark.sql.DataFrame** = []

hive> select \* from student;

OK

1 Song M 39

2 Jay F 29

3 Roger M 60

4 Tim M 27

Time taken: 0.854 seconds, Fetched: 4 row(s)

[11] HDFS

linsong@linsongdeMacBook-Air ~ % **cd /Users/linsong/Documents/2021/BigData/bigdata**

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -mkdir -p /dagger/pool2021**

2021-01-27 22:56:25,886 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -ls /**

2021-01-27 22:56:34,377 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 3 items

drwxr-xr-x - linsong supergroup 0 2021-01-27 22:56 /dagger

drwx-wx-wx - linsong supergroup 0 2021-01-26 23:32 /tmp

drwxr-xr-x - linsong supergroup 0 2021-01-26 23:32 /user

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -put /Users/linsong/Documents/2021/BigData/bigdata/resources/pool.txt /dagger/pool2021**

2021-01-27 22:59:53,153 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -cat /dagger/pool2021/pool.txt**

2021-01-27 23:01:56,406 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

1 Song 39

2 Jay 29

3 Roger 60**%**

linsong@linsongdeMacBook-Air bigdata % h**dfs dfs -get /dagger/pool2021/pool.txt /Users/linsong/Documents/2021/BigData/bigdata/download/poolDup.txt**

2021-01-27 23:09:33,817 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -rm /dagger/pool2021/pool.txt**

2021-01-27 23:10:46,408 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Deleted /dagger/pool2021/pool.txt

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -rm -r /dagger/**

2021-01-27 23:11:03,614 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Deleted /dagger

linsong@linsongdeMacBook-Air bigdata % **hdfs dfs -ls /**

2021-01-27 23:11:22,422 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 2 items

drwx-wx-wx - linsong supergroup 0 2021-01-26 23:32 /tmp

drwxr-xr-x - linsong supergroup 0 2021-01-26 23:32 /user

[12] Misc

HDFS browser: <http://localhost:9870/explorer.html#/>

http://192.168.11.21:8080/

[13] PySpark

> pip install pyspark

linsong@linsongdeMacBook-Air ~ % pyspark

>>> from pyspark import SparkContext

>>> from pyspark import SparkFiles

>>> from \_\_future\_\_ import print\_function

>>> rdd = sc.textFile("/Users/linsong/Documents/2021/BigData/bigdata/download/poolDup.txt")

>>> result = rdd.flatMap(lambda line: line.split(" ")).map(lambda line: (line, 1)).reduceByKey(lambda a, b: a + b)

>>> result.foreach(print);

[Stage 0:> (0 + 2) / 2]/usr/local/Cellar/apache-spark/2.4.5/libexec/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling

/usr/local/Cellar/apache-spark/2.4.5/libexec/python/lib/pyspark.zip/pyspark/shuffle.py:60: UserWarning: Please install psutil to have better support with spilling

(u'1', 1)

(u'60', 1)

(u'39', 1)

(u'Roger', 1)

(u'3', 1)

(u'2', 1)

(u'Song', 1)

(u'29', 1)

(u'Jay', 1)