Hadoop Environment Ready for Developer Basic Verification

▼ TOC

Overview

因為 Hadoop Cluster 乃分散式系統，配置上較為複雜，需要處理相依性。為確認 Hadoop 叢集環境元件已經正確配置完成，因此在開發前，需請系統建置者執行元件測試；以確保測試通過，就可以交接給開發者使用。

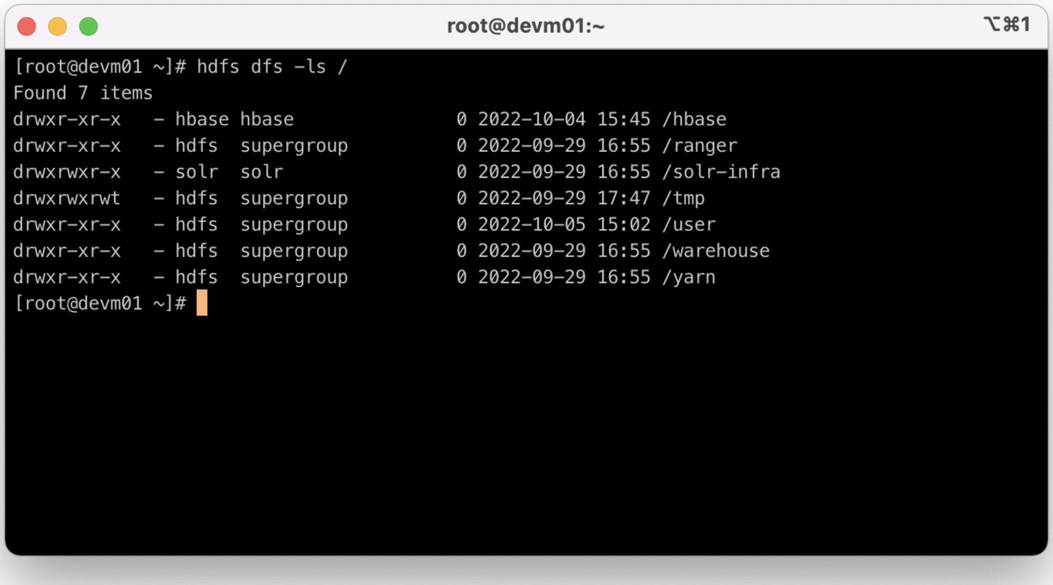
Components of Hadoop Ecosystem

以下截圖皆為 Cloudera Hadoop (CDP) 之執行結果畫面，僅供參考。

▼ HDFS

- 查看資料夾

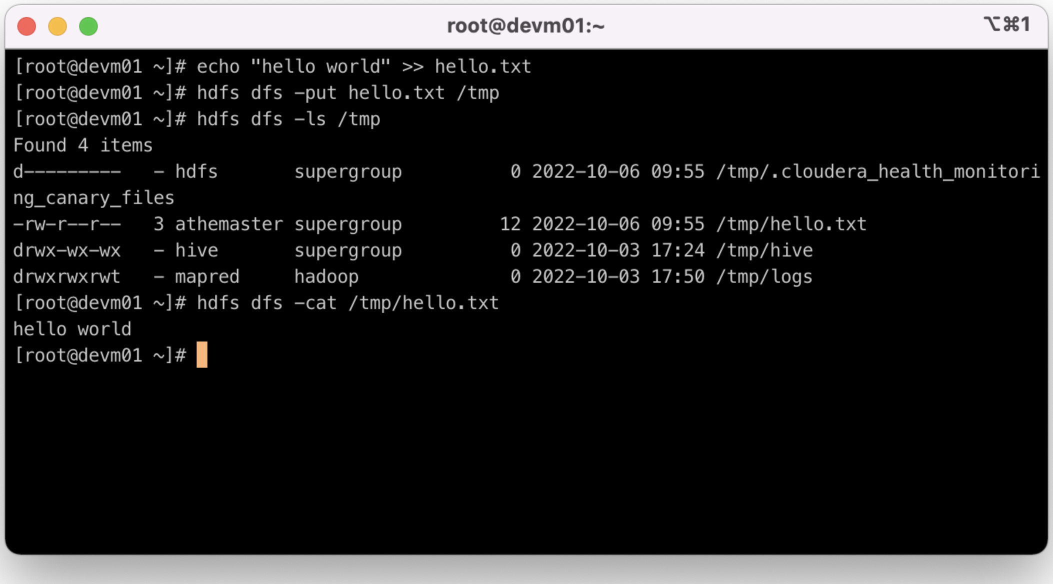
```
# 查看資料夾
hdfs dfs -ls /
```



```
[root@devm01 ~]# hdfs dfs -ls /
Found 7 items
drwxr-xr-x   - hbase hbase               0 2022-10-04 15:45 /hbase
drwxr-xr-x   - hdfs  supergroup          0 2022-09-29 16:55 /ranger
drwxrwxr-x   - solr  solr                0 2022-09-29 16:55 /solr-infra
drwxrwxrwt   - hdfs  supergroup          0 2022-09-29 17:47 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2022-10-05 15:02 /user
drwxr-xr-x   - hdfs  supergroup          0 2022-09-29 16:55 /warehouse
drwxr-xr-x   - hdfs  supergroup          0 2022-09-29 16:55 /yarn
[root@devm01 ~]# 
```

- 測試丟入檔案

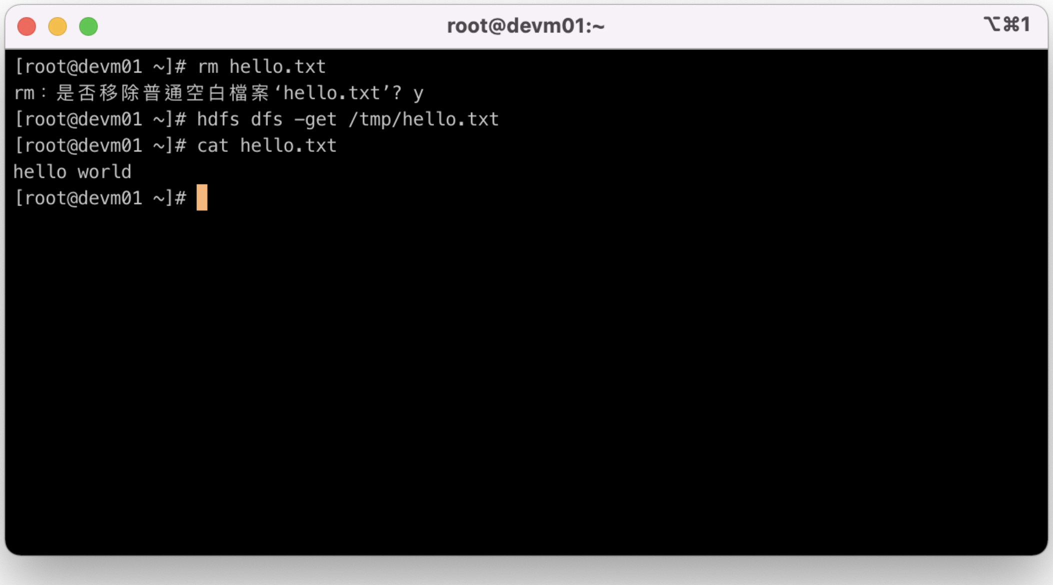
```
echo "hello world" >> hello.txt
hdfs dfs -put hello.txt /tmp
hdfs dfs -ls /tmp
hdfs dfs -cat /tmp/hello.txt
```



```
[root@devm01 ~]# echo "hello world" >> hello.txt
[root@devm01 ~]# hdfs dfs -put hello.txt /tmp
[root@devm01 ~]# hdfs dfs -ls /tmp
Found 4 items
d---------   - hdfs       supergroup          0 2022-10-06 09:55 /tmp/.cloudera_health_monitori
ng_canary_files
-rw-r--r--   3 athemaster supergroup         12 2022-10-06 09:55 /tmp/hello.txt
drwx-wx-wx   - hive       supergroup          0 2022-10-03 17:24 /tmp/hive
drwxrwxrwt   - mapred     hadoop              0 2022-10-03 17:50 /tmp/logs
[root@devm01 ~]# hdfs dfs -cat /tmp/hello.txt
hello world
[root@devm01 ~]# 
```

- 取出檔案

```
rm hello.txt
hdfs dfs -get /tmp/hello.txt
cat hello.txt
```



```
[root@devm01 ~]# rm hello.txt
rm：是否移除普通空白檔案‘hello.txt’? y
[root@devm01 ~]# hdfs dfs -get /tmp/hello.txt
[root@devm01 ~]# cat hello.txt
hello world
[root@devm01 ~]#
```

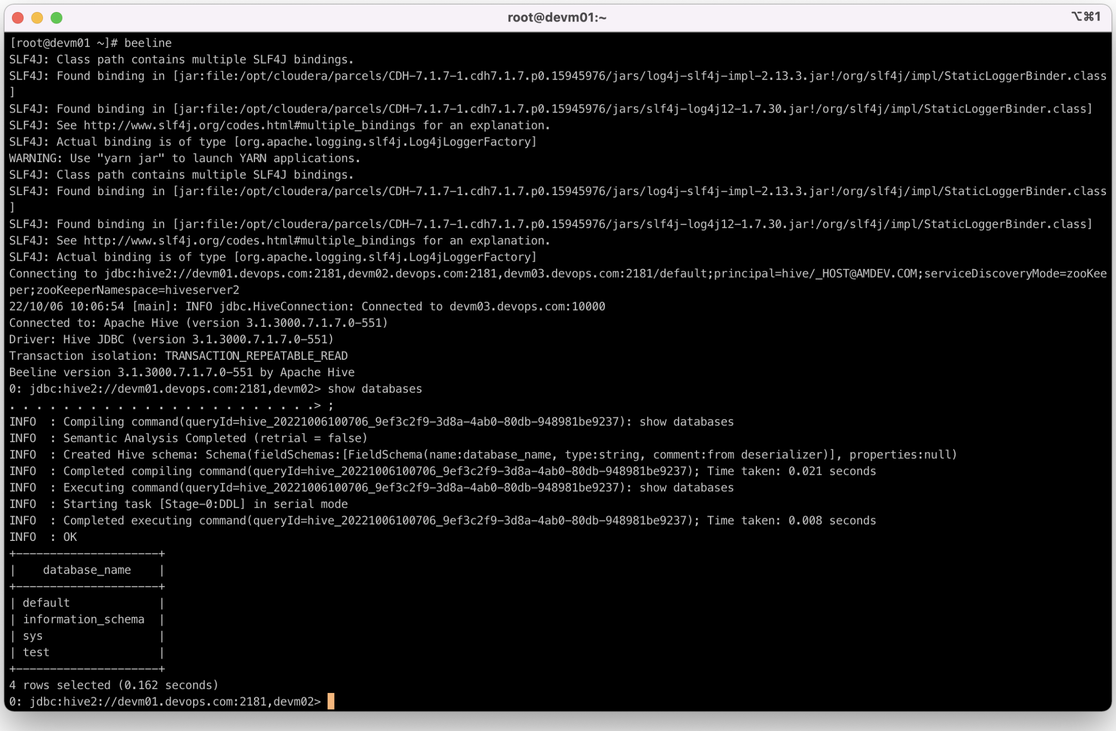
▼ Hive

- beeline

```
beeline
show databases

create external table if not exists default.users (
  name string,
  age int
);

insert into default.users values
 ("Alice", 16),  ("Bob", 18),
 ("Cindy", 53),  ("Joe", 28),
 ("Apple", 5),   ("John", 33),
 ("Roger", 45),  ("Kevin", 17),
 ("Gordon", 30), ("Paul", 33),
 ("Sandy", 39),  ("Alex", 70);

create table if not exists default.users_internal (
  name string,
  age int
);

insert into default.users_internal values
 ("Alice", 16),  ("Bob", 18),
 ("Cindy", 53),  ("Joe", 28),
 ("Apple", 5),   ("John", 33),
 ("Roger", 45),  ("Kevin", 17),
 ("Gordon", 30), ("Paul", 33),
 ("Sandy", 39),  ("Alex", 70);
```

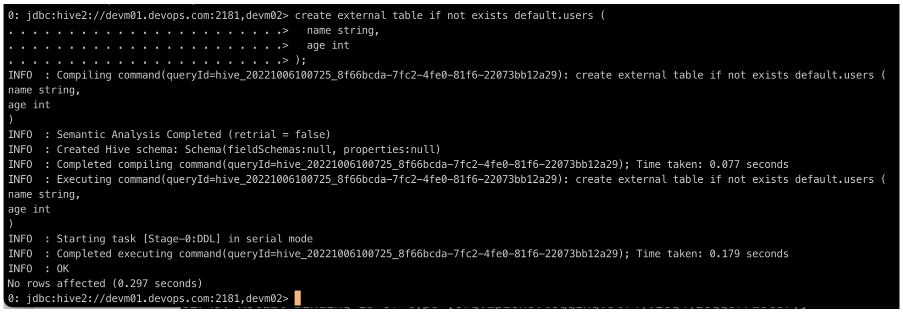
```
select * from default.users;
select * from default.users_internal;
```

show databases;



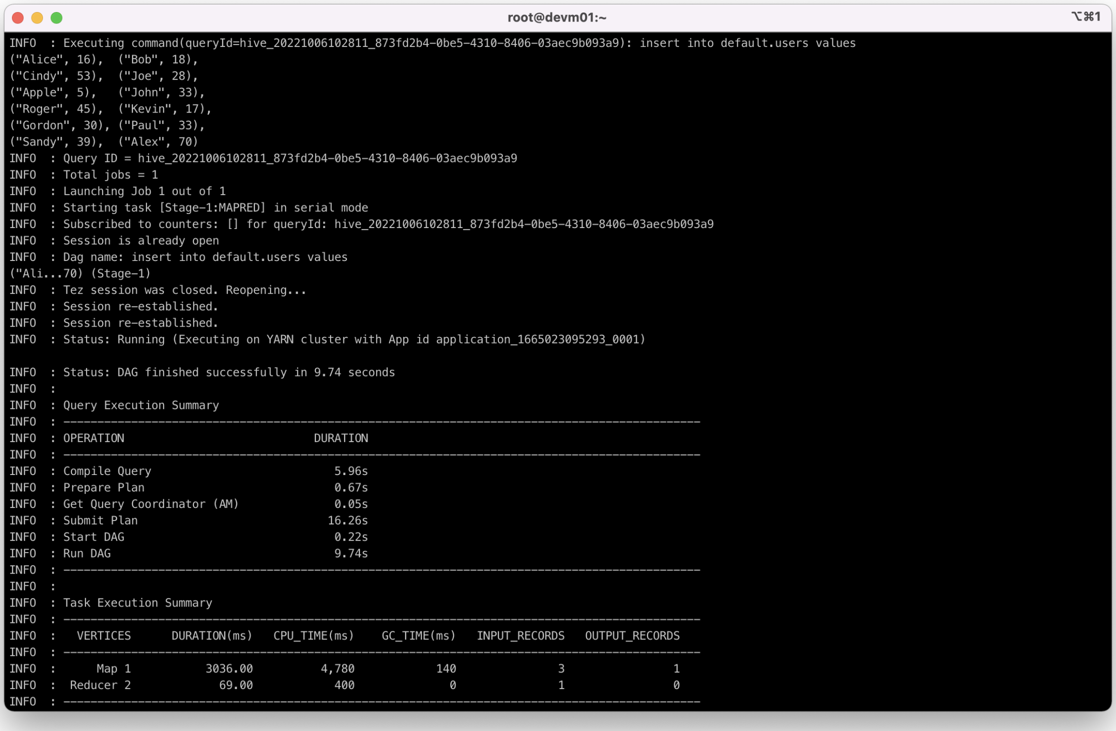
```
[root@devm01 ~]# beeline
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/log4j-slf4j-impl-2.13.3.jar!/org/slf4j/impl/StaticLoggerBinder.class
]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/slf4j-log4j12-1.7.30.jar!/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.apache.logging.slf4j.Log4jLoggerFactory]
WARNING: Use "yarn jar" to launch YARN applications.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/log4j-slf4j-impl-2.13.3.jar!/org/slf4j/impl/StaticLoggerBinder.class
]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/slf4j-log4j12-1.7.30.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.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://devm01.devops.com:2181,devm02.devops.com:2181,devm03.devops.com:2181/default;principal=hive/_HOST@AMDEV.COM;serviceDiscoveryMode=zooKee
per;zooKeeperNamespace=hiveserver2
22/10/06 10:06:54 [main]: INFO jdbc.HiveConnection: Connected to devm03.devops.com:10000
Connected to: Apache Hive (version 3.1.3000.7.1.7.0-551)
Driver: Hive JDBC (version 3.1.3000.7.1.7.0-551)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3000.7.1.7.0-551 by Apache Hive
0: jdbc:hive2://devm01.devops.com:2181,devm02> show databases
. . . . . . . . . . . . . . . . . . . . . . .> ;
INFO  : Compiling command(queryId=hive_20221006100706_9ef3c2f9-3d8a-4ab0-80db-948981be9237): show databases
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20221006100706_9ef3c2f9-3d8a-4ab0-80db-948981be9237); Time taken: 0.021 seconds
INFO  : Executing command(queryId=hive_20221006100706_9ef3c2f9-3d8a-4ab0-80db-948981be9237): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221006100706_9ef3c2f9-3d8a-4ab0-80db-948981be9237); Time taken: 0.008 seconds
INFO  : OK
+---------------------+
|    database_name    |
+---------------------+
| default             |
| information_schema  |
| sys                 |
| test                |
+---------------------+
4 rows selected (0.162 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

建立 external table



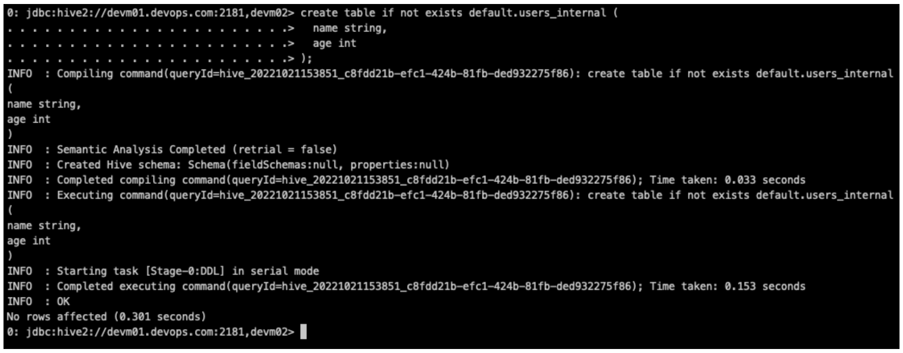
```
0: jdbc:hive2://devm01.devops.com:2181,devm02> create external table if not exists default.users (
. . . . . . . . . . . . . . . . . . . . . . .>   name string,
. . . . . . . . . . . . . . . . . . . . . . .>   age int
. . . . . . . . . . . . . . . . . . . . . . .> );
INFO  : Compiling command(queryId=hive_20221006100725_8f66bcda-7fc2-4fe0-81f6-22073bb12a29): create external table if not exists default.users (
name string,
age int
)
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20221006100725_8f66bcda-7fc2-4fe0-81f6-22073bb12a29); Time taken: 0.077 seconds
INFO  : Executing command(queryId=hive_20221006100725_8f66bcda-7fc2-4fe0-81f6-22073bb12a29): create external table if not exists default.users (
name string,
age int
)
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221006100725_8f66bcda-7fc2-4fe0-81f6-22073bb12a29); Time taken: 0.179 seconds
INFO  : OK
No rows affected (0.297 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

寫入 external table



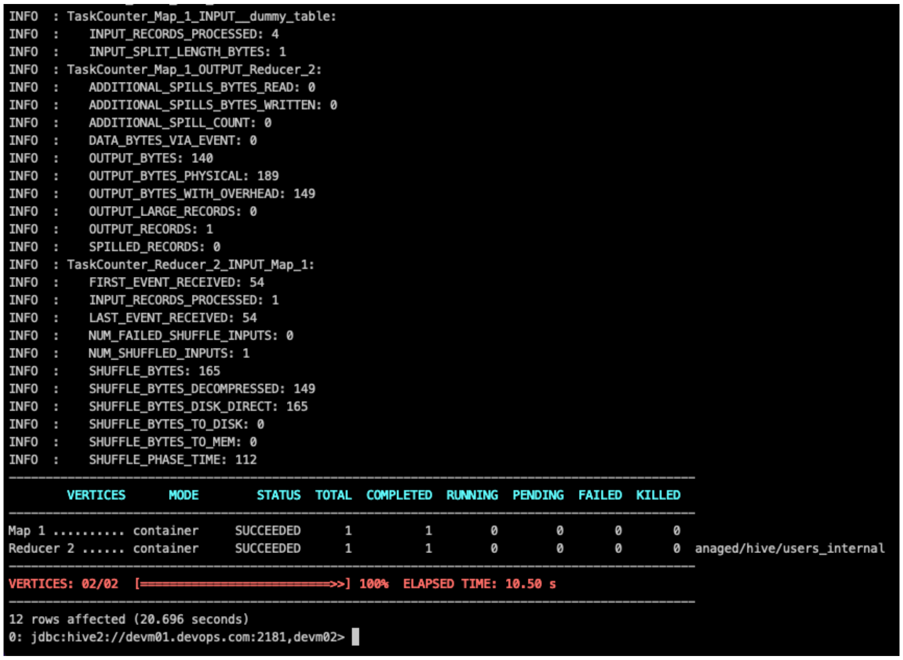
```
INFO  : Executing command(queryId=hive_20221006102811_873fd2b4-0be5-4310-8406-03aec9b093a9): insert into default.users values
("Alice", 16),  ("Bob", 18),
("Cindy", 53),  ("Joe", 28),
("Apple", 5),   ("John", 33),
("Roger", 45),  ("Kevin", 17),
("Gordon", 30), ("Paul", 33),
("Sandy", 39),  ("Alex", 70)
INFO  : Query ID = hive_20221006102811_873fd2b4-0be5-4310-8406-03aec9b093a9
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20221006102811_873fd2b4-0be5-4310-8406-03aec9b093a9
INFO  : Session is already open
INFO  : Dag name: insert into default.users values
("Ali...70) (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1665023095293_0001)

INFO  : Status: DAG finished successfully in 9.74 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                           5.96s
INFO  : Prepare Plan                            0.67s
INFO  : Get Query Coordinator (AM)              0.05s
INFO  : Submit Plan                            16.26s
INFO  : Start DAG                               0.22s
INFO  : Run DAG                                 9.74s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)   CPU_TIME(ms)    GC_TIME(ms)   INPUT_RECORDS    OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :      Map 1           3036.00          4,780            140               3                1
INFO  :  Reducer 2             69.00            400              0               1                0
INFO  : ----------------------------------------------------------------------------------------------
```

建立 internal_table



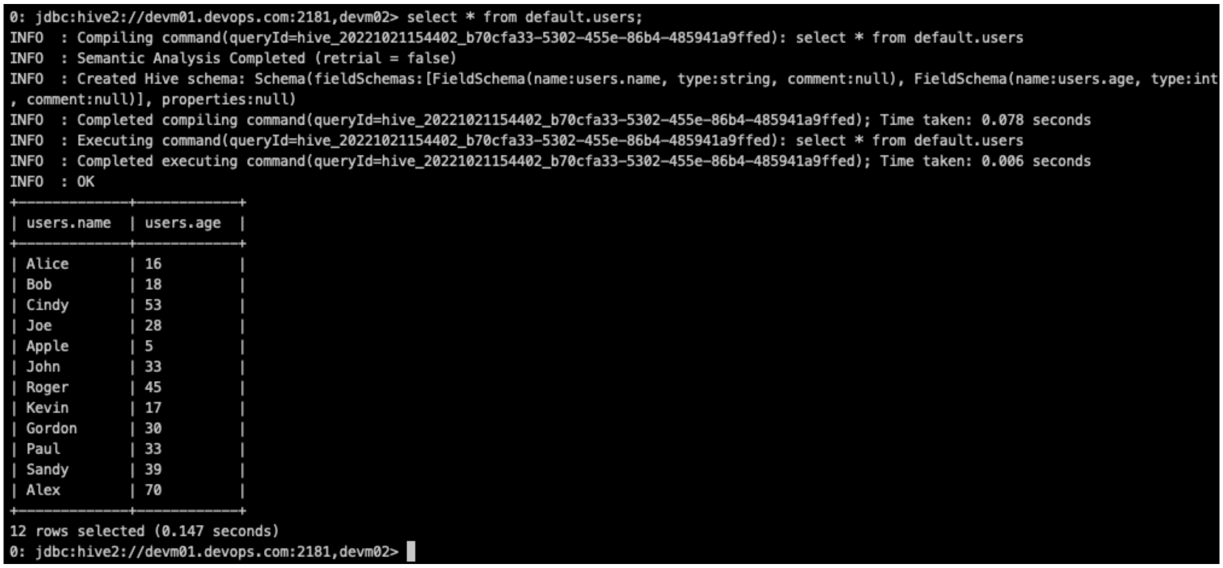
```
0: jdbc:hive2://devm01.devops.com:2181,devm02> create table if not exists default.users_internal (
. . . . . . . . . . . . . . . . . . . . . . .>   name string,
. . . . . . . . . . . . . . . . . . . . . . .>   age int
. . . . . . . . . . . . . . . . . . . . . . .> );
INFO  : Compiling command(queryId=hive_20221021153851_c8fdd21b-efc1-424b-81fb-ded932275f86): create table if not exists default.users_internal
(
name string,
age int
)
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20221021153851_c8fdd21b-efc1-424b-81fb-ded932275f86); Time taken: 0.033 seconds
INFO  : Executing command(queryId=hive_20221021153851_c8fdd21b-efc1-424b-81fb-ded932275f86): create table if not exists default.users_internal
(
name string,
age int
)
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221021153851_c8fdd21b-efc1-424b-81fb-ded932275f86); Time taken: 0.153 seconds
INFO  : OK
No rows affected (0.301 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

寫入 internal_table



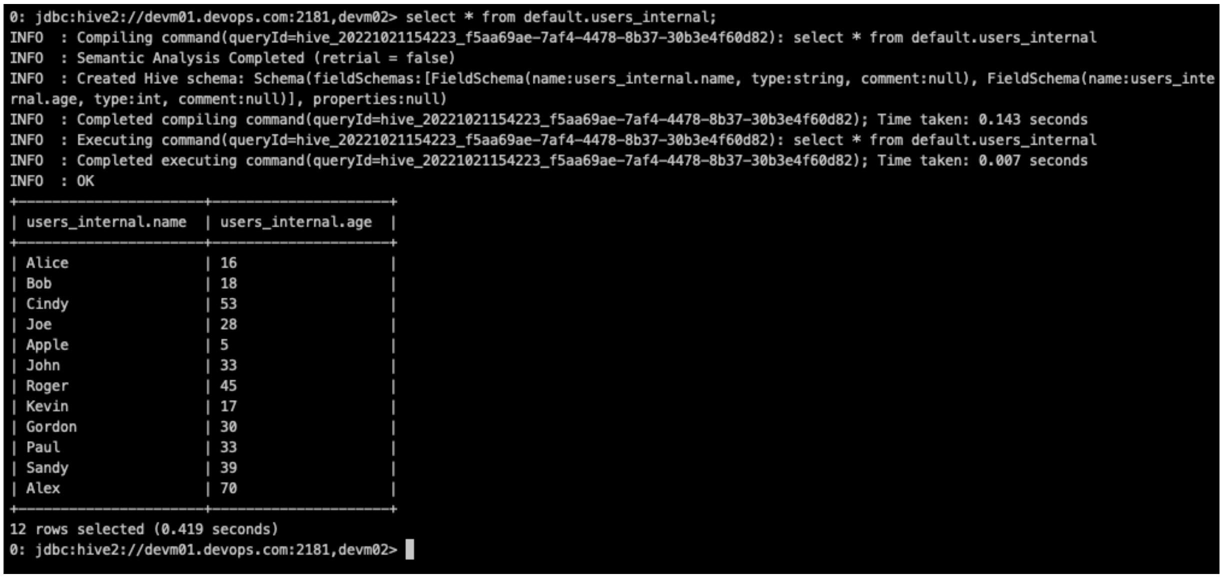
```
INFO  : TaskCounter_Map_1_INPUT__dummy_table:
INFO  :    INPUT_RECORDS_PROCESSED: 4
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 1
INFO  : TaskCounter_Map_1_OUTPUT_Reducer_2:
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 0
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    ADDITIONAL_SPILL_COUNT: 0
INFO  :    DATA_BYTES_VIA_EVENT: 0
INFO  :    OUTPUT_BYTES: 140
INFO  :    OUTPUT_BYTES_PHYSICAL: 189
INFO  :    OUTPUT_BYTES_WITH_OVERHEAD: 149
INFO  :    OUTPUT_LARGE_RECORDS: 0
INFO  :    OUTPUT_RECORDS: 1
INFO  :    SPILLED_RECORDS: 0
INFO  : TaskCounter_Reducer_2_INPUT_Map_1:
INFO  :    FIRST_EVENT_RECEIVED: 54
INFO  :    INPUT_RECORDS_PROCESSED: 1
INFO  :    LAST_EVENT_RECEIVED: 54
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    NUM_SHUFFLED_INPUTS: 1
INFO  :    SHUFFLE_BYTES: 165
INFO  :    SHUFFLE_BYTES_DECOMPRESSED: 149
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 165
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_PHASE_TIME: 112
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0  anaged/hive/users_internal
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.50 s
----------------------------------------------------------------------------------------------
12 rows affected (20.696 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

查詢 external table



```
0: jdbc:hive2://devm01.devops.com:2181,devm02> select * from default.users;
INFO  : Compiling command(queryId=hive_20221021154402_b70cfa33-5302-455e-86b4-485941a9ffed): select * from default.users
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:users.name, type:string, comment:null), FieldSchema(name:users.age, type:int
, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20221021154402_b70cfa33-5302-455e-86b4-485941a9ffed); Time taken: 0.078 seconds
INFO  : Executing command(queryId=hive_20221021154402_b70cfa33-5302-455e-86b4-485941a9ffed): select * from default.users
INFO  : Completed executing command(queryId=hive_20221021154402_b70cfa33-5302-455e-86b4-485941a9ffed); Time taken: 0.006 seconds
INFO  : OK
+-------------+------------+
| users.name  | users.age  |
+-------------+------------+
| Alice       | 16         |
| Bob         | 18         |
| Cindy       | 53         |
| Joe         | 28         |
| Apple       | 5          |
| John        | 33         |
| Roger       | 45         |
| Kevin       | 17         |
| Gordon      | 30         |
| Paul        | 33         |
| Sandy       | 39         |
| Alex        | 70         |
+-------------+------------+
12 rows selected (0.147 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

查詢 internal table

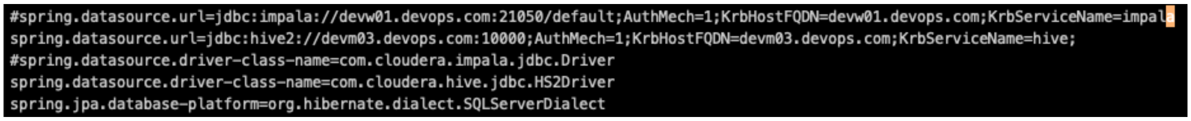


```
0: jdbc:hive2://devm01.devops.com:2181,devm02> select * from default.users_internal;
INFO  : Compiling command(queryId=hive_20221021154223_f5aa69ae-7af4-4478-8b37-30b3e4f60d82): select * from default.users_internal
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:users_internal.name, type:string, comment:null), FieldSchema(name:users_inte
rnal.age, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20221021154223_f5aa69ae-7af4-4478-8b37-30b3e4f60d82); Time taken: 0.143 seconds
INFO  : Executing command(queryId=hive_20221021154223_f5aa69ae-7af4-4478-8b37-30b3e4f60d82): select * from default.users_internal
INFO  : Completed executing command(queryId=hive_20221021154223_f5aa69ae-7af4-4478-8b37-30b3e4f60d82); Time taken: 0.007 seconds
INFO  : OK
+----------------------+---------------------+
| users_internal.name  | users_internal.age  |
+----------------------+---------------------+
| Alice                | 16                  |
| Bob                  | 18                  |
| Cindy                | 53                  |
| Joe                  | 28                  |
| Apple                | 5                   |
| John                 | 33                  |
| Roger                | 45                  |
| Kevin                | 17                  |
| Gordon               | 30                  |
| Paul                 | 33                  |
| Sandy                | 39                  |
| Alex                 | 70                  |
+----------------------+---------------------+
12 rows selected (0.419 seconds)
0: jdbc:hive2://devm01.devops.com:2181,devm02>
```

- JDBC

使用 jdbc-kerberos-0.0.1-SNAPSHOT.jar 進行測試

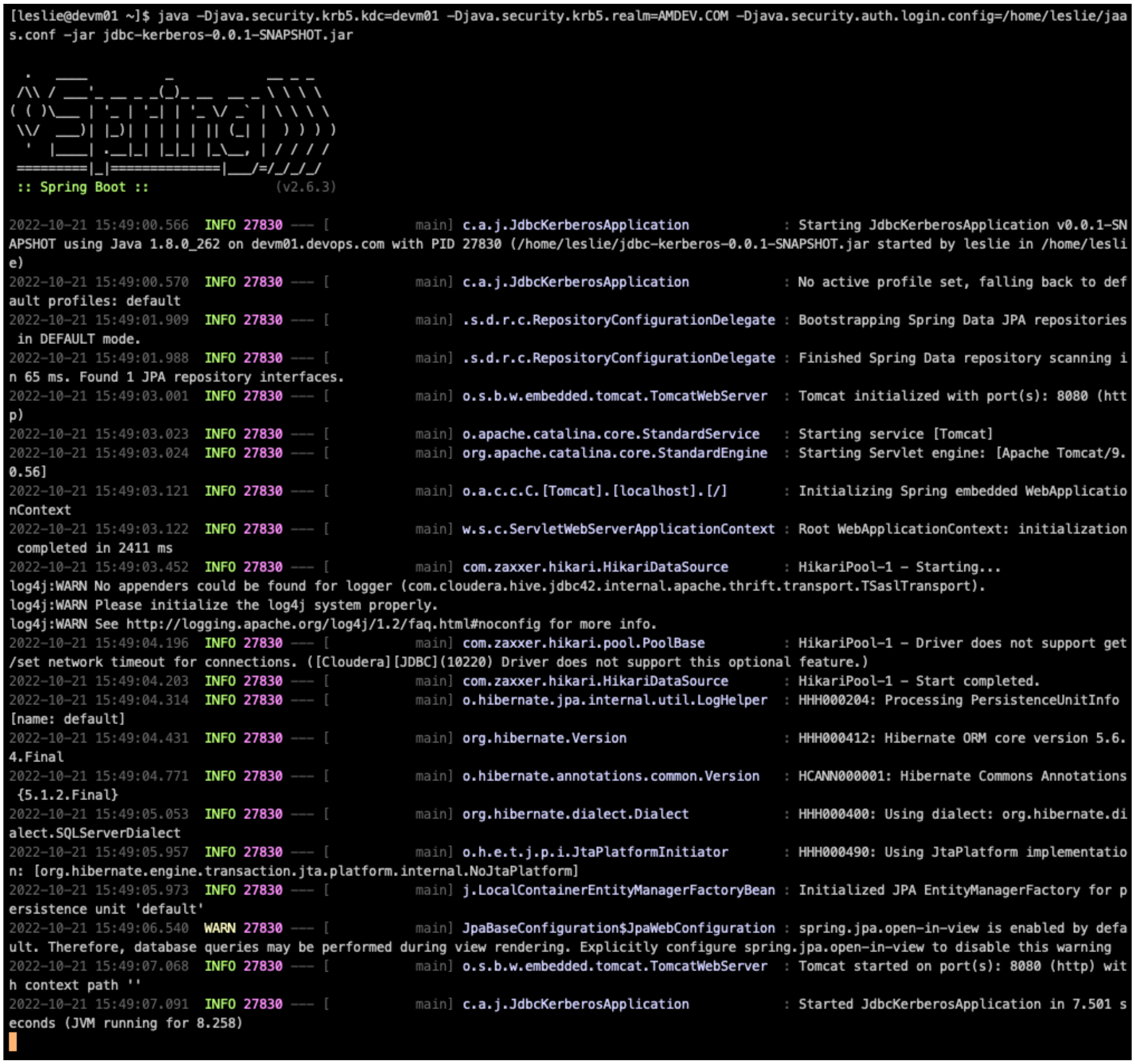
- 先確認 application.properties 使用 hive 的設定 (圖片僅供參考)



```
#spring.datasource.url=jdbc:impala://devw01.devops.com:21050/default;AuthMech=1;KrbHostFQDN=devw01.devops.com;KrbServiceName=impala
spring.datasource.url=jdbc:hive2://devm03.devops.com:10000;AuthMech=1;KrbHostFQDN=devm03.devops.com;KrbServiceName=hive;
#spring.datasource.driver-class-name=com.cloudera.impala.jdbc.Driver
spring.datasource.driver-class-name=com.cloudera.hive.jdbc.HS2Driver
spring.jpa.database-platform=org.hibernate.dialect.SQLServerDialect
```

- 啟動 web 程式 (圖片啟動指令僅供參考)

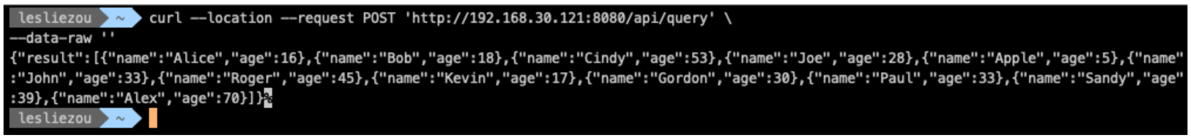
```
java -jar jdbc-kerberos-0.0.1-SNAPSHOT.jar
```



```
[leslie@devm01 ~]$ java -Djava.security.krb5.kdc=devm01 -Djava.security.krb5.realm=AMDEV.COM -Djava.security.auth.login.config=/home/leslie/jaa
s.conf -jar jdbc-kerberos-0.0.1-SNAPSHOT.jar

  .   ____          _            __ _ _
 /\\ / ___'_ __ _ _(_)_ __  __ _ \ \ \ \
( ( )\___ | '_ | '_| | '_ \/ _` | \ \ \ \
 \\/  ___)| |_)| | | | | || (_| |  ) ) ) )
  '  |____| .__|_| |_|_| |_\__, | / / / /
 =========|_|==============|___/=/_/_/_/
 :: Spring Boot ::                (v2.6.3)

2022-10-21 15:49:00.566  INFO 27830 --- [           main] c.a.j.JdbcKerberosApplication            : Starting JdbcKerberosApplication v0.0.1-SN
APSHOT using Java 1.8.0_262 on devm01.devops.com with PID 27830 (/home/leslie/jdbc-kerberos-0.0.1-SNAPSHOT.jar started by leslie in /home/lesli
e)
2022-10-21 15:49:00.570  INFO 27830 --- [           main] c.a.j.JdbcKerberosApplication            : No active profile set, falling back to def
ault profiles: default
2022-10-21 15:49:01.909  INFO 27830 --- [           main] .s.d.r.c.RepositoryConfigurationDelegate : Bootstrapping Spring Data JPA repositories
 in DEFAULT mode.
2022-10-21 15:49:01.988  INFO 27830 --- [           main] .s.d.r.c.RepositoryConfigurationDelegate : Finished Spring Data repository scanning i
n 65 ms. Found 1 JPA repository interfaces.
2022-10-21 15:49:03.001  INFO 27830 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat initialized with port(s): 8080 (htt
p)
2022-10-21 15:49:03.023  INFO 27830 --- [           main] o.apache.catalina.core.StandardService   : Starting service [Tomcat]
2022-10-21 15:49:03.024  INFO 27830 --- [           main] org.apache.catalina.core.StandardEngine  : Starting Servlet engine: [Apache Tomcat/9.
0.56]
2022-10-21 15:49:03.121  INFO 27830 --- [           main] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring embedded WebApplicatio
nContext
2022-10-21 15:49:03.122  INFO 27830 --- [           main] w.s.c.ServletWebServerApplicationContext : Root WebApplicationContext: initialization
 completed in 2411 ms
2022-10-21 15:49:03.452  INFO 27830 --- [           main] com.zaxxer.hikari.HikariDataSource       : HikariPool-1 - Starting...
log4j:WARN No appenders could be found for logger (com.cloudera.hive.jdbc42.internal.apache.thrift.transport.TSaslTransport).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2022-10-21 15:49:04.196  INFO 27830 --- [           main] com.zaxxer.hikari.pool.PoolBase          : HikariPool-1 - Driver does not support get
/set network timeout for connections. ([Cloudera][JDBC](10220) Driver does not support this optional feature.)
2022-10-21 15:49:04.203  INFO 27830 --- [           main] com.zaxxer.hikari.HikariDataSource       : HikariPool-1 - Start completed.
2022-10-21 15:49:04.314  INFO 27830 --- [           main] o.hibernate.jpa.internal.util.LogHelper  : HHH000204: Processing PersistenceUnitInfo
[name: default]
2022-10-21 15:49:04.431  INFO 27830 --- [           main] org.hibernate.Version                    : HHH000412: Hibernate ORM core version 5.6.
4.Final
2022-10-21 15:49:04.771  INFO 27830 --- [           main] o.hibernate.annotations.common.Version   : HCANN000001: Hibernate Commons Annotations
 {5.1.2.Final}
2022-10-21 15:49:05.053  INFO 27830 --- [           main] org.hibernate.dialect.Dialect            : HHH000400: Using dialect: org.hibernate.di
alect.SQLServerDialect
2022-10-21 15:49:05.957  INFO 27830 --- [           main] o.h.e.t.j.p.i.JtaPlatformInitiator       : HHH000490: Using JtaPlatform implementatio
n: [org.hibernate.engine.transaction.jta.platform.internal.NoJtaPlatform]
2022-10-21 15:49:05.973  INFO 27830 --- [           main] j.LocalContainerEntityManagerFactoryBean : Initialized JPA EntityManagerFactory for p
ersistence unit 'default'
2022-10-21 15:49:06.540  WARN 27830 --- [           main] JpaBaseConfiguration$JpaWebConfiguration : spring.jpa.open-in-view is enabled by defa
ult. Therefore, database queries may be performed during view rendering. Explicitly configure spring.jpa.open-in-view to disable this warning
2022-10-21 15:49:07.068  INFO 27830 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat started on port(s): 8080 (http) wit
h context path ''
2022-10-21 15:49:07.091  INFO 27830 --- [           main] c.a.j.JdbcKerberosApplication            : Started JdbcKerberosApplication in 7.501 s
econds (JVM running for 8.258)
```

- 另開 Terminal， 透過 CURL 指令呼叫 (圖片啟動指令僅供參考)

```
curl --location --request POST 'http://<ip>:8080/api/query' \
--data-raw ''
```

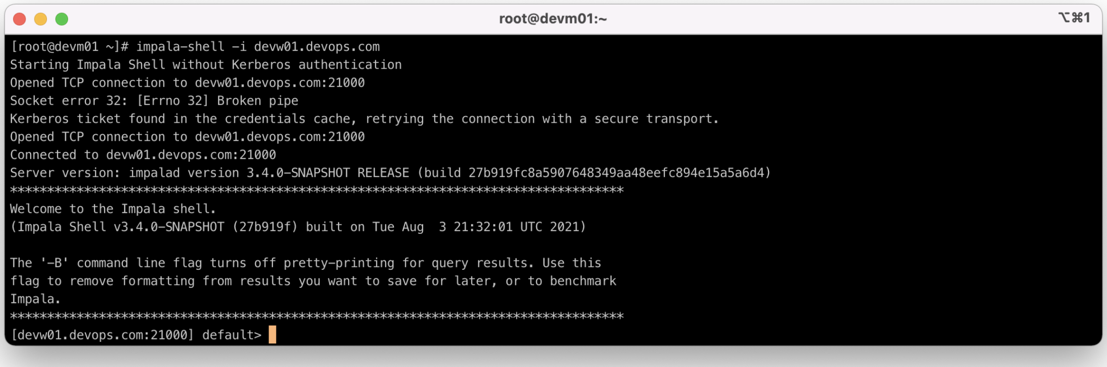


```
lesliezou  ~  curl --location --request POST 'http://192.168.30.121:8080/api/query' \
--data-raw ''
{"result":[{"name":"Alice","age":16},{"name":"Bob","age":18},{"name":"Cindy","age":53},{"name":"Joe","age":28},{"name":"Apple","age":5},{"name"
:"John","age":33},{"name":"Roger","age":45},{"name":"Kevin","age":17},{"name":"Gordon","age":30},{"name":"Paul","age":33},{"name":"Sandy","age"
:39},{"name":"Alex","age":70}]}%
lesliezou  ~
```

Impala

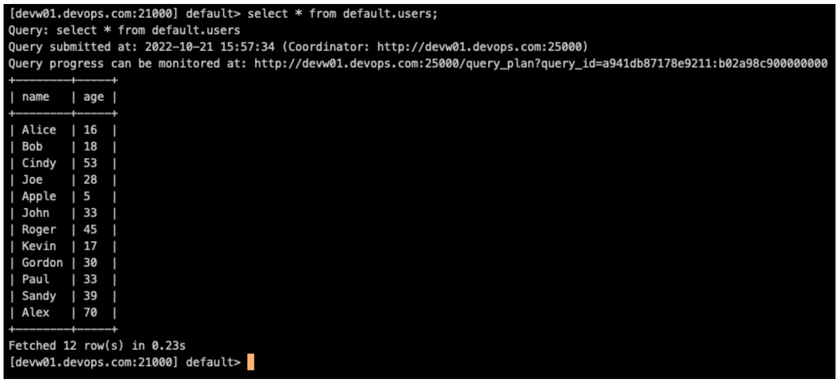
- Impala-shell (圖片僅供參考)

```
impala-shell -i <host_name>

select * from default.users;
select * from default.users_internal;
```



```
[root@devm01 ~]# impala-shell -i devw01.devops.com
Starting Impala Shell without Kerberos authentication
Opened TCP connection to devw01.devops.com:21000
Socket error 32: [Errno 32] Broken pipe
Kerberos ticket found in the credentials cache, retrying the connection with a secure transport.
Opened TCP connection to devw01.devops.com:21000
Connected to devw01.devops.com:21000
Server version: impalad version 3.4.0-SNAPSHOT RELEASE (build 27b919fc8a5907648349aa48eefc894e15a5a6d4)
***********************************************************************************
Welcome to the Impala shell.
(Impala Shell v3.4.0-SNAPSHOT (27b919f) built on Tue Aug  3 21:32:01 UTC 2021)

The '-B' command line flag turns off pretty-printing for query results. Use this
flag to remove formatting from results you want to save for later, or to benchmark
Impala.
***********************************************************************************
[devw01.devops.com:21000] default>
```

查詢 external table

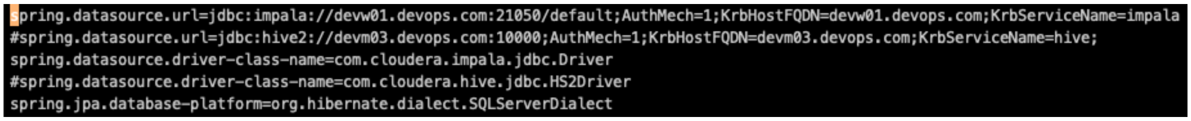


```
[devw01.devops.com:21000] default> select * from default.users;
Query: select * from default.users
Query submitted at: 2022-10-21 15:57:34 (Coordinator: http://devw01.devops.com:25000)
Query progress can be monitored at: http://devw01.devops.com:25000/query_plan?query_id=a941db87178e9211:b02a98c900000000
+--------+-----+
| name   | age |
+--------+-----+
| Alice  | 16  |
| Bob    | 18  |
| Cindy  | 53  |
| Joe    | 28  |
| Apple  | 5   |
| John   | 33  |
| Roger  | 45  |
| Kevin  | 17  |
| Gordon | 30  |
| Paul   | 33  |
| Sandy  | 39  |
| Alex   | 70  |
+--------+-----+
Fetched 12 row(s) in 0.23s
[devw01.devops.com:21000] default>
```

- JDBC

使用 jdbc-kerberos-0.0.1-SNAPSHOT.jar 進行測試

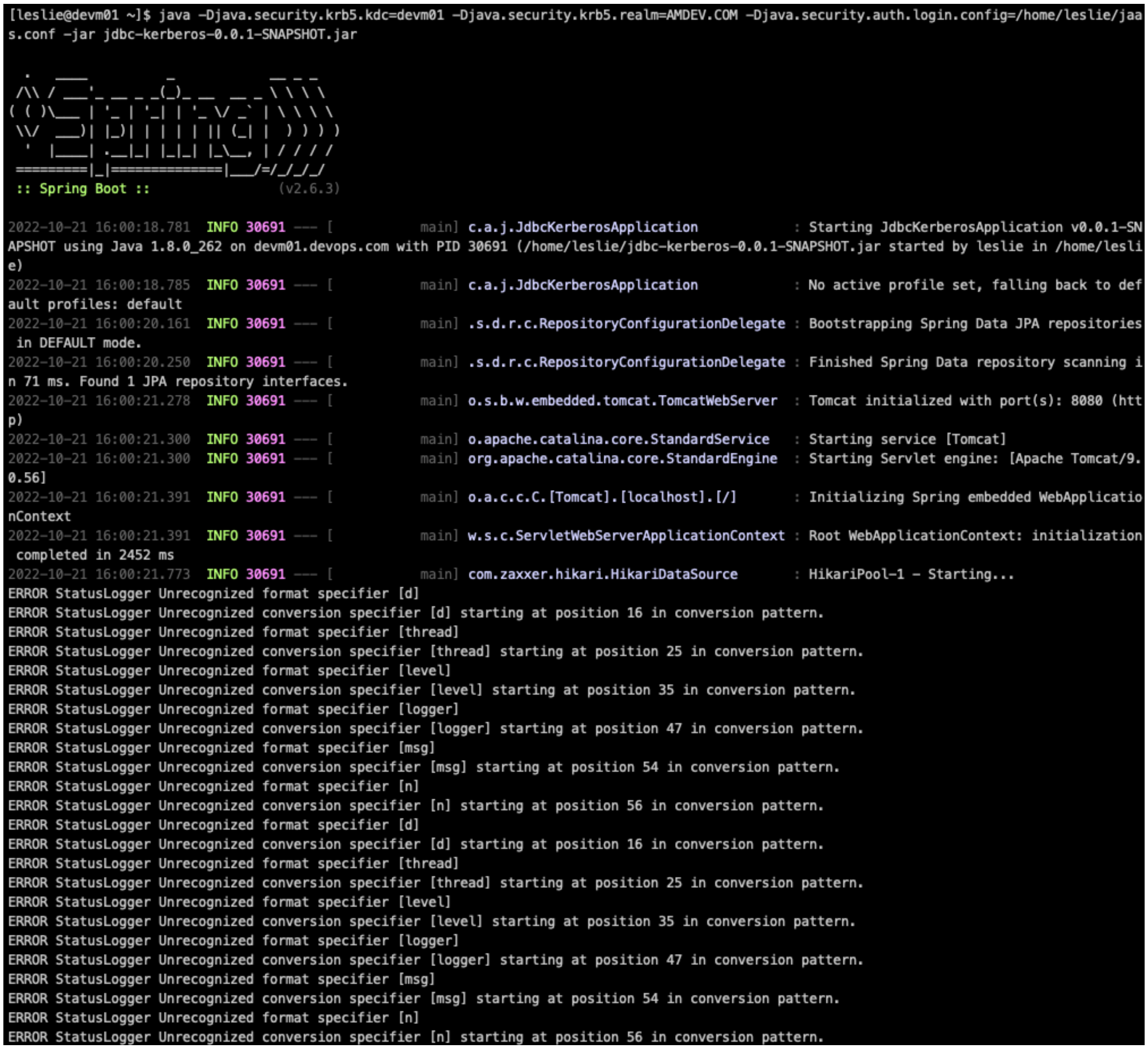
- 先確認 application.properties 使用 impala 的設定 (圖片僅供參考)



```
spring.datasource.url=jdbc:impala://devw01.devops.com:21050/default;AuthMech=1;KrbHostFQDN=devw01.devops.com;KrbServiceName=impala
#spring.datasource.url=jdbc:hive2://devm03.devops.com:10000;AuthMech=1;KrbHostFQDN=devm03.devops.com;KrbServiceName=hive;
spring.datasource.driver-class-name=com.cloudera.impala.jdbc.Driver
#spring.datasource.driver-class-name=com.cloudera.hive.jdbc.HS2Driver
spring.jpa.database-platform=org.hibernate.dialect.SQLServerDialect
```

- 啟動 web 程式 (圖片啟動指令僅供參考, log 過多提供部分截圖)

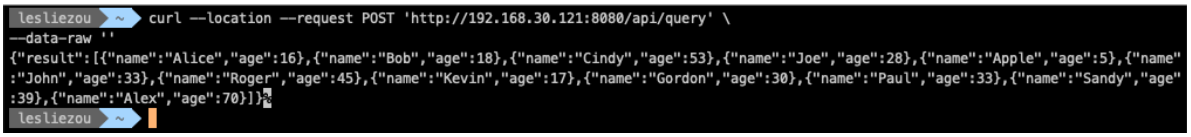
```
java -jar jdbc-kerberos-0.0.1-SNAPSHOT.jar
```



```
[leslie@devm01 ~]$ java -Djava.security.krb5.kdc=devm01 -Djava.security.krb5.realm=AMDEV.COM -Djava.security.auth.login.config=/home/leslie/jaa
s.conf -jar jdbc-kerberos-0.0.1-SNAPSHOT.jar

  .   ____          _            __ _ _
 /\\ / ___'_ __ _ _(_)_ __  __ _ \ \ \ \
( ( )\___ | '_ | '_| | '_ \/ _` | \ \ \ \
 \\/  ___)| |_)| | | | | || (_| |  ) ) ) )
  '  |____| .__|_| |_|_| |_\__, | / / / /
 =========|_|==============|___/=/_/_/_/
 :: Spring Boot ::                (v2.6.3)

2022-10-21 16:00:18.781  INFO 30691 --- [           main] c.a.j.JdbcKerberosApplication            : Starting JdbcKerberosApplication v0.0.1-SN
APSHOT using Java 1.8.0_262 on devm01.devops.com with PID 30691 (/home/leslie/jdbc-kerberos-0.0.1-SNAPSHOT.jar started by leslie in /home/lesli
e)
2022-10-21 16:00:18.785  INFO 30691 --- [           main] c.a.j.JdbcKerberosApplication            : No active profile set, falling back to def
ault profiles: default
2022-10-21 16:00:20.161  INFO 30691 --- [           main] .s.d.r.c.RepositoryConfigurationDelegate : Bootstrapping Spring Data JPA repositories
 in DEFAULT mode.
2022-10-21 16:00:20.250  INFO 30691 --- [           main] .s.d.r.c.RepositoryConfigurationDelegate : Finished Spring Data repository scanning i
n 71 ms. Found 1 JPA repository interfaces.
2022-10-21 16:00:21.278  INFO 30691 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat initialized with port(s): 8080 (htt
p)
2022-10-21 16:00:21.300  INFO 30691 --- [           main] o.apache.catalina.core.StandardService   : Starting service [Tomcat]
2022-10-21 16:00:21.300  INFO 30691 --- [           main] org.apache.catalina.core.StandardEngine  : Starting Servlet engine: [Apache Tomcat/9.
0.56]
2022-10-21 16:00:21.391  INFO 30691 --- [           main] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring embedded WebApplicatio
nContext
2022-10-21 16:00:21.391  INFO 30691 --- [           main] w.s.c.ServletWebServerApplicationContext : Root WebApplicationContext: initialization
 completed in 2452 ms
2022-10-21 16:00:21.773  INFO 30691 --- [           main] com.zaxxer.hikari.HikariDataSource       : HikariPool-1 - Starting...
ERROR StatusLogger Unrecognized format specifier [d]
ERROR StatusLogger Unrecognized conversion specifier [d] starting at position 16 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [thread]
ERROR StatusLogger Unrecognized conversion specifier [thread] starting at position 25 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [level]
ERROR StatusLogger Unrecognized conversion specifier [level] starting at position 35 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [logger]
ERROR StatusLogger Unrecognized conversion specifier [logger] starting at position 47 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [msg]
ERROR StatusLogger Unrecognized conversion specifier [msg] starting at position 54 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [n]
ERROR StatusLogger Unrecognized conversion specifier [n] starting at position 56 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [d]
ERROR StatusLogger Unrecognized conversion specifier [d] starting at position 16 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [thread]
ERROR StatusLogger Unrecognized conversion specifier [thread] starting at position 25 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [level]
ERROR StatusLogger Unrecognized conversion specifier [level] starting at position 35 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [logger]
ERROR StatusLogger Unrecognized conversion specifier [logger] starting at position 47 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [msg]
ERROR StatusLogger Unrecognized conversion specifier [msg] starting at position 54 in conversion pattern.
ERROR StatusLogger Unrecognized format specifier [n]
ERROR StatusLogger Unrecognized conversion specifier [n] starting at position 56 in conversion pattern.
```

- 另開 Terminal，透過 CURL 指令呼叫 (圖片啟動指令僅供參考)

```
curl --location --request POST 'http://<ip>:8080/api/query' \
--data-raw ''
```



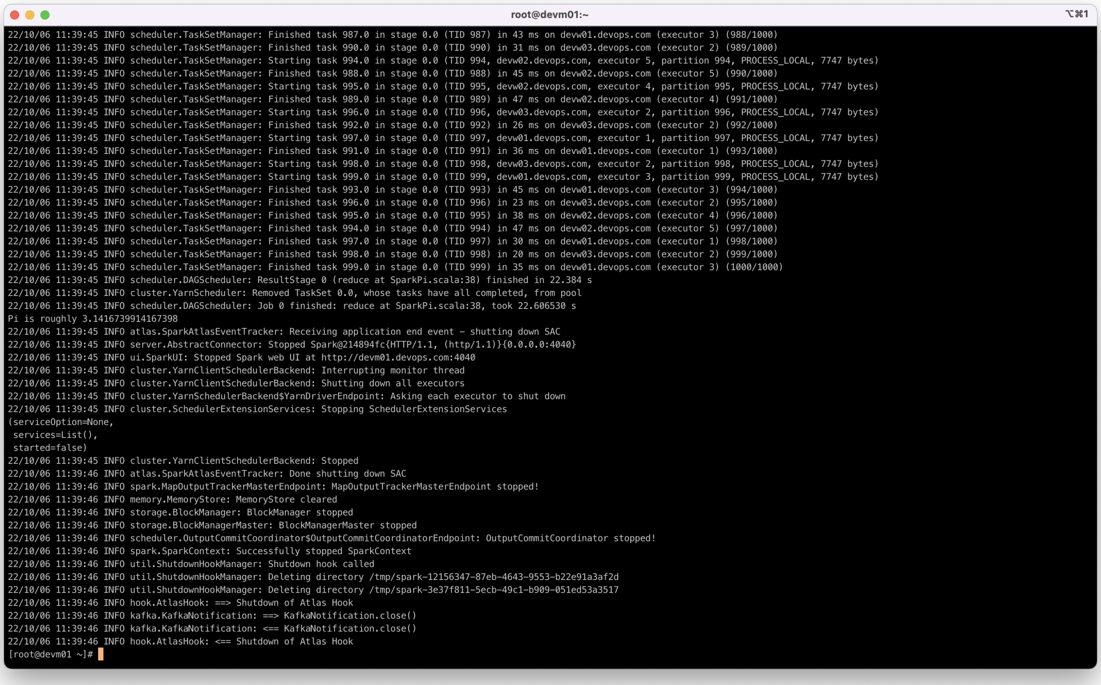
```
 lesliezou  ~  curl --location --request POST 'http://192.168.30.121:8080/api/query' \
--data-raw ''
{"result":[{"name":"Alice","age":16},{"name":"Bob","age":18},{"name":"Cindy","age":53},{"name":"Joe","age":28},{"name":"Apple","age":5},{"name"
:"John","age":33},{"name":"Roger","age":45},{"name":"Kevin","age":17},{"name":"Gordon","age":30},{"name":"Paul","age":33},{"name":"Sandy","age"
:39},{"name":"Alex","age":70}]}%
 lesliezou  ~ 
```

▼ Spark

▼ Scala

- 執行計算 pi (jar 檔路徑僅供參考，可能會因為版本有差異)

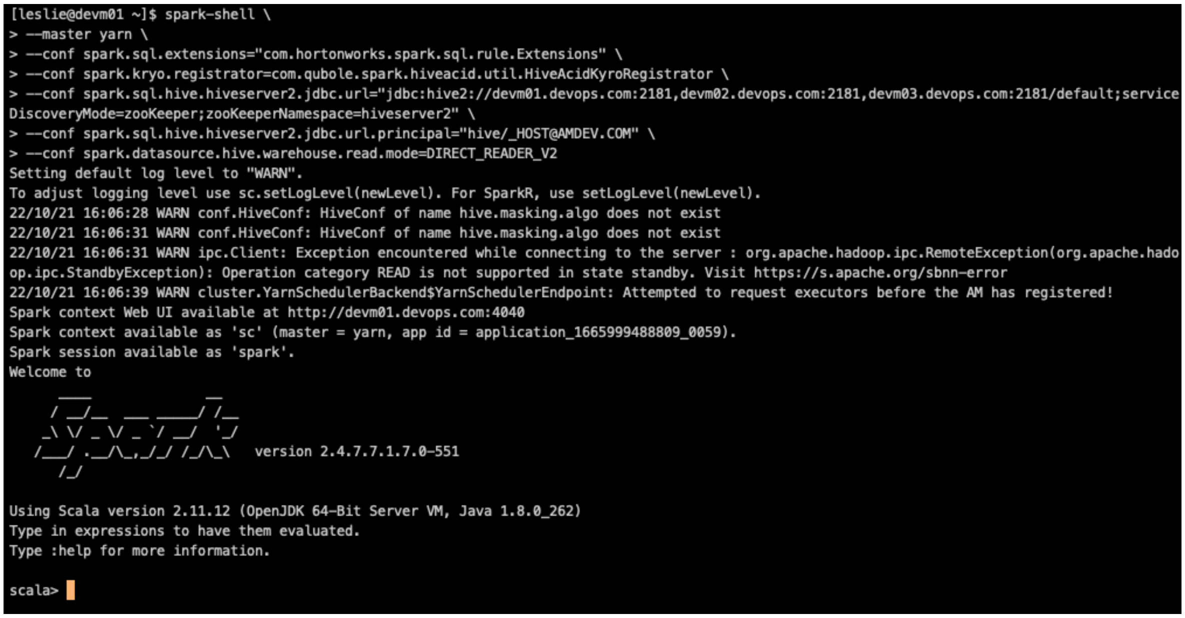
```
spark-submit --class org.apache.spark.examples.SparkPi \
 --master yarn-client \
 /opt/cloudera/parcels/CDH/lib/spark/examples/jars/spark-examples_2.11-2.4.7.7.1.7.0-551.jar 1000
```



```
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 987.0 in stage 0.0 (TID 987) in 43 ms on devw01.devops.com (executor 3) (988/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 990.0 in stage 0.0 (TID 990) in 31 ms on devw03.devops.com (executor 2) (989/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 994.0 in stage 0.0 (TID 994, devw02.devops.com, executor 5, partition 994, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 988.0 in stage 0.0 (TID 988) in 45 ms on devw02.devops.com (executor 5) (990/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 995.0 in stage 0.0 (TID 995, devw02.devops.com, executor 4, partition 995, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 989.0 in stage 0.0 (TID 989) in 47 ms on devw02.devops.com (executor 4) (991/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 996.0 in stage 0.0 (TID 996, devw03.devops.com, executor 2, partition 996, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 992.0 in stage 0.0 (TID 992) in 26 ms on devw03.devops.com (executor 2) (992/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 997.0 in stage 0.0 (TID 997, devw01.devops.com, executor 1, partition 997, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 991.0 in stage 0.0 (TID 991) in 36 ms on devw01.devops.com (executor 1) (993/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 998.0 in stage 0.0 (TID 998, devw03.devops.com, executor 2, partition 998, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Starting task 999.0 in stage 0.0 (TID 999, devw01.devops.com, executor 3, partition 999, PROCESS_LOCAL, 7747 bytes)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 993.0 in stage 0.0 (TID 993) in 45 ms on devw01.devops.com (executor 3) (994/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 996.0 in stage 0.0 (TID 996) in 23 ms on devw03.devops.com (executor 2) (995/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 995.0 in stage 0.0 (TID 995) in 38 ms on devw02.devops.com (executor 4) (996/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 994.0 in stage 0.0 (TID 994) in 47 ms on devw02.devops.com (executor 5) (997/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 997.0 in stage 0.0 (TID 997) in 30 ms on devw01.devops.com (executor 1) (998/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 998.0 in stage 0.0 (TID 998) in 20 ms on devw03.devops.com (executor 2) (999/1000)
22/10/06 11:39:45 INFO scheduler.TaskSetManager: Finished task 999.0 in stage 0.0 (TID 999) in 35 ms on devw01.devops.com (executor 3) (1000/1000)
22/10/06 11:39:45 INFO scheduler.DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38) finished in 22.384 s
22/10/06 11:39:45 INFO cluster.YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
22/10/06 11:39:45 INFO scheduler.DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 22.606530 s
Pi is roughly 3.1416739914167398
22/10/06 11:39:45 INFO atlas.SparkAtlasEventTracker: Receiving application end event - shutting down SAC
22/10/06 11:39:45 INFO server.AbstractConnector: Stopped Spark@214894fc{HTTP/1.1, (http/1.1)}{0.0.0.0:4040}
22/10/06 11:39:45 INFO ui.SparkUI: Stopped Spark web UI at http://devm01.devops.com:4040
22/10/06 11:39:45 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread
22/10/06 11:39:45 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
22/10/06 11:39:45 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/10/06 11:39:45 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/10/06 11:39:45 INFO cluster.YarnClientSchedulerBackend: Stopped
22/10/06 11:39:46 INFO atlas.SparkAtlasEventTracker: Done shutting down SAC
22/10/06 11:39:46 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/10/06 11:39:46 INFO memory.MemoryStore: MemoryStore cleared
22/10/06 11:39:46 INFO storage.BlockManager: BlockManager stopped
22/10/06 11:39:46 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
22/10/06 11:39:46 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/10/06 11:39:46 INFO spark.SparkContext: Successfully stopped SparkContext
22/10/06 11:39:46 INFO util.ShutdownHookManager: Shutdown hook called
22/10/06 11:39:46 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-12156347-87eb-4643-9553-b22e91a3af2d
22/10/06 11:39:46 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-3e37f811-5ecb-49c1-b909-051ed53a3517
22/10/06 11:39:46 INFO hook.AtlasHook: ==> Shutdown of Atlas Hook
22/10/06 11:39:46 INFO kafka.KafkaNotification: ==> KafkaNotification.close()
22/10/06 11:39:46 INFO kafka.KafkaNotification: <== KafkaNotification.close()
22/10/06 11:39:46 INFO hook.AtlasHook: <== Shutdown of Atlas Hook
[root@devm01 ~]#
```

- 使用 HWC (指令僅供參考，可能會因為版本有差異)

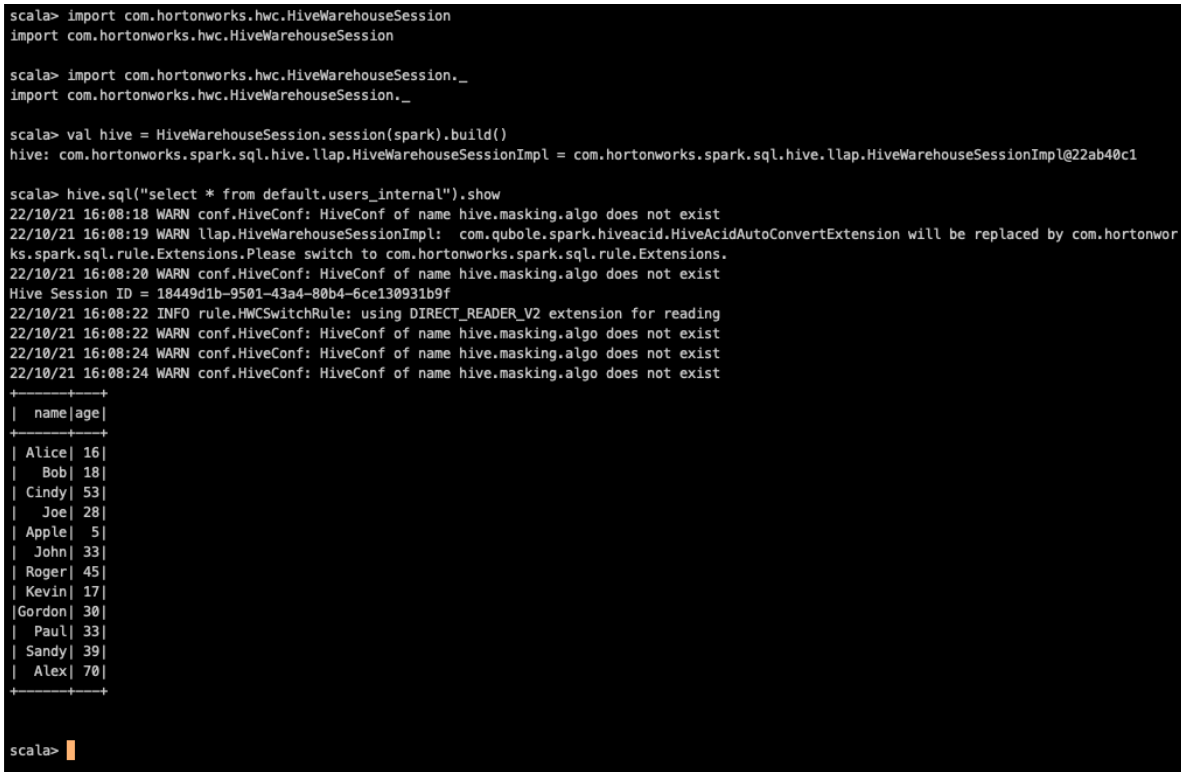
```
spark-shell \
--master yarn \
--conf spark.sql.extensions="com.hortonworks.spark.sql.rule.Extensions" \
--conf spark.kryo.registrator=com.qubole.spark.hiveacid.util.HiveAcidKyroRegistrator \
--conf spark.sql.hive.hiveserver2.jdbc.url="jdbc:hive2://devm01.devops.com:2181,devm02.devops.com:2181,devm03.devops.com:2181/defau
--conf spark.sql.hive.hiveserver2.jdbc.url.principal="hive/_HOST@AMDEV.COM" \
--conf spark.datasource.hive.warehouse.read.mode=DIRECT_READER_V2
```



```
[leslie@devm01 ~]$ spark-shell \
> --master yarn \
> --conf spark.sql.extensions="com.hortonworks.spark.sql.rule.Extensions" \
> --conf spark.kryo.registrator=com.qubole.spark.hiveacid.util.HiveAcidKyroRegistrator \
> --conf spark.sql.hive.hiveserver2.jdbc.url="jdbc:hive2://devm01.devops.com:2181,devm02.devops.com:2181,devm03.devops.com:2181/default;service
DiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2" \
> --conf spark.sql.hive.hiveserver2.jdbc.url.principal="hive/_HOST@AMDEV.COM" \
> --conf spark.datasource.hive.warehouse.read.mode=DIRECT_READER_V2
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/10/21 16:06:28 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:06:31 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:06:31 WARN ipc.Client: Exception encountered while connecting to the server : org.apache.hadoop.ipc.RemoteException(org.apache.hado
op.ipc.StandbyException): Operation category READ is not supported in state standby. Visit https://s.apache.org/sbnn-error
22/10/21 16:06:39 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
Spark context Web UI available at http://devm01.devops.com:4040
Spark context available as 'sc' (master = yarn, app id = application_1665999488809_0059).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.7.7.1.7.0-551
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_262)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- 使用 HWC 查詢 Internal table 資料

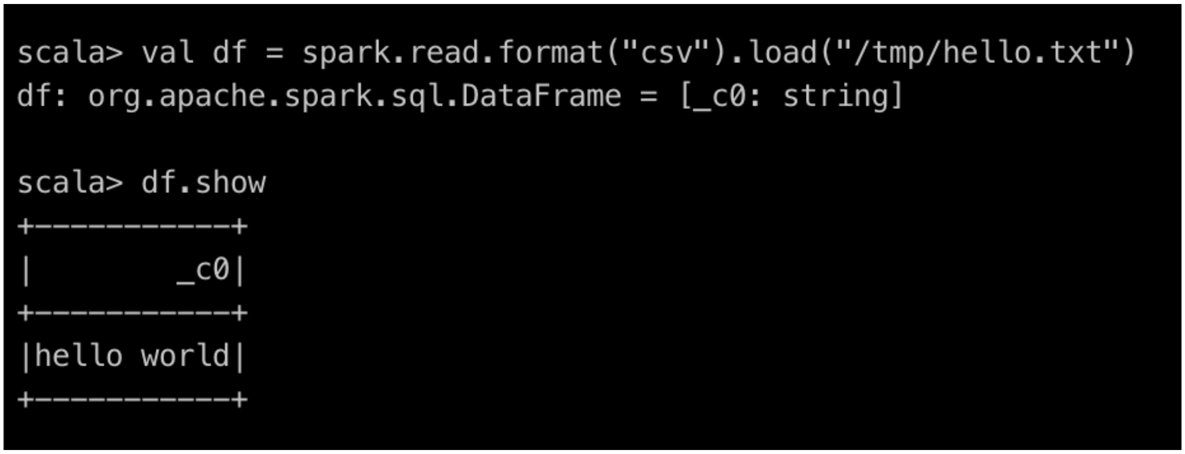
```
import com.hortonworks.hwc.HiveWarehouseSession
import com.hortonworks.hwc.HiveWarehouseSession._
val hive = HiveWarehouseSession.session(spark).build()
hive.sql("select * from default.users_internal").show
```



```
scala> import com.hortonworks.hwc.HiveWarehouseSession
import com.hortonworks.hwc.HiveWarehouseSession

scala> import com.hortonworks.hwc.HiveWarehouseSession._
import com.hortonworks.hwc.HiveWarehouseSession._

scala> val hive = HiveWarehouseSession.session(spark).build()
hive: com.hortonworks.spark.sql.hive.llap.HiveWarehouseSessionImpl = com.hortonworks.spark.sql.hive.llap.HiveWarehouseSessionImpl@22ab40c1

scala> hive.sql("select * from default.users_internal").show
22/10/21 16:08:18 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:08:19 WARN llap.HiveWarehouseSessionImpl:  com.qubole.spark.hiveacid.HiveAcidAutoConvertExtension will be replaced by com.hortonwor
ks.spark.sql.rule.Extensions.Please switch to com.hortonworks.spark.sql.rule.Extensions.
22/10/21 16:08:20 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
Hive Session ID = 18449d1b-9501-43a4-80b4-6ce130931b9f
22/10/21 16:08:22 INFO rule.HWCSwitchRule: using DIRECT_READER_V2 extension for reading
22/10/21 16:08:22 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:08:24 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:08:24 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
+------+---+
|  name|age|
+------+---+
| Alice| 16|
|   Bob| 18|
| Cindy| 53|
|   Joe| 28|
| Apple|  5|
|  John| 33|
| Roger| 45|
| Kevin| 17|
|Gordon| 30|
|  Paul| 33|
| Sandy| 39|
|  Alex| 70|
+------+---+


scala> 
```

- 讀取資料 (HDFS)

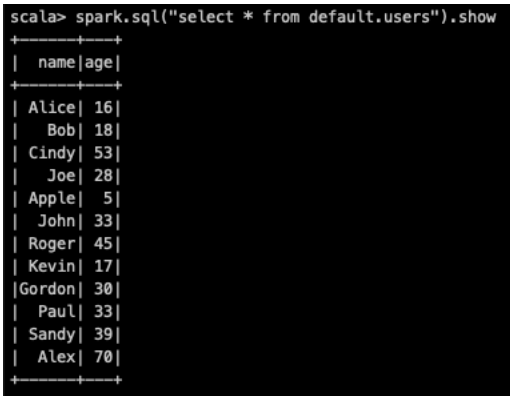
```
val df = spark.read.format("csv").load("/tmp/hello.txt")
df.show
```



```
scala> val df = spark.read.format("csv").load("/tmp/hello.txt")
df: org.apache.spark.sql.DataFrame = [_c0: string]

scala> df.show
+-----------+
|        _c0|
+-----------+
|hello world|
+-----------+
```

- 讀取資料 (External table)

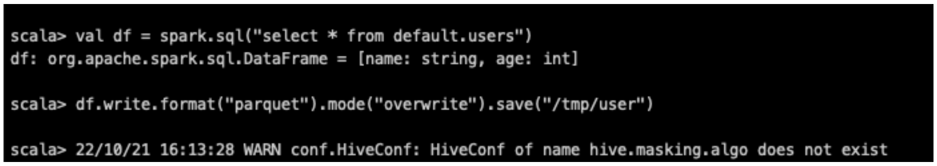
```
spark.sql("select * from default.users").show
```



```
scala> spark.sql("select * from default.users").show
+------+---+
|  name|age|
+------+---+
| Alice| 16|
|   Bob| 18|
| Cindy| 53|
|   Joe| 28|
| Apple|  5|
|  John| 33|
| Roger| 45|
| Kevin| 17|
|Gordon| 30|
|  Paul| 33|
| Sandy| 39|
|  Alex| 70|
+------+---+
```

- 寫入資料至 HDFS

```
val df = spark.sql("select * from default.users")
df.write.format("parquet").mode("overwrite").save("/tmp/user")
```

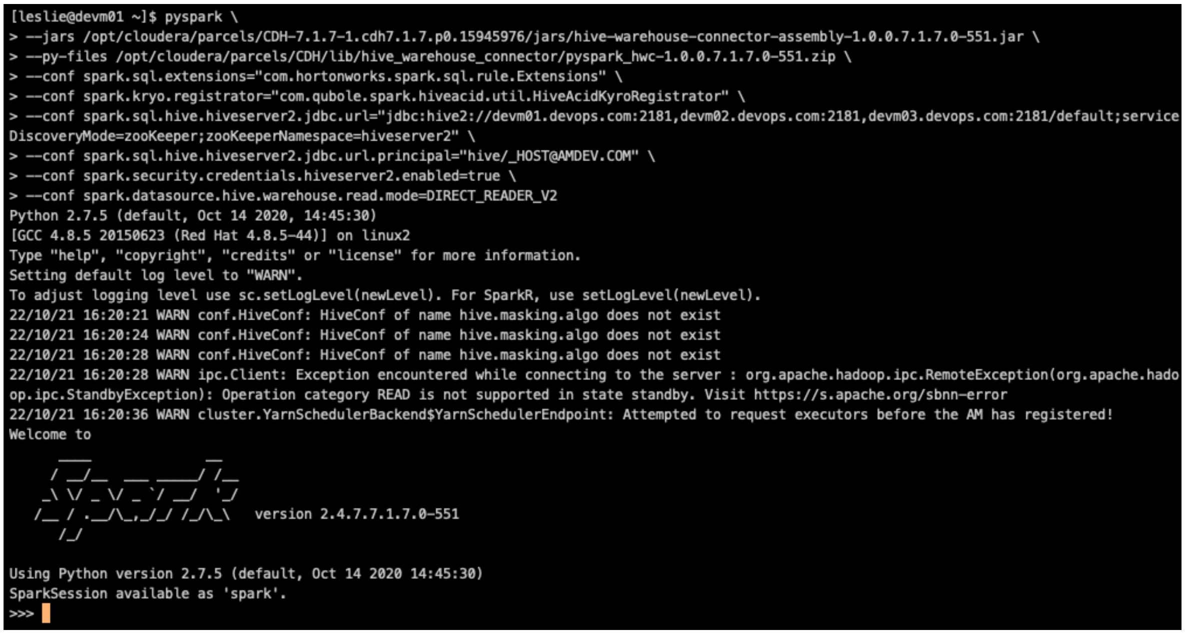


```
scala> val df = spark.sql("select * from default.users")
df: org.apache.spark.sql.DataFrame = [name: string, age: int]

scala> df.write.format("parquet").mode("overwrite").save("/tmp/user")

scala> 22/10/21 16:13:28 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
```

▼ PySpark

- 使用 HWC (指令僅供參考，可能會因為版本有差異)

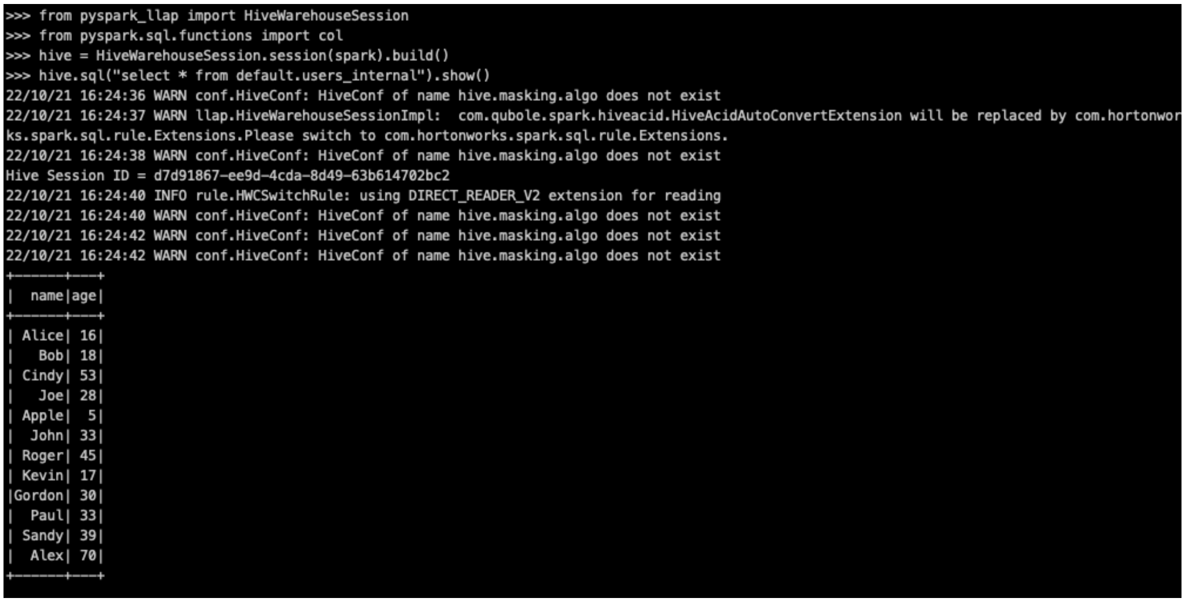
```
pyspark \
--jars /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hive-warehouse-connector-assembly-1.0.0.7.1.7.0-551.jar \
--py-files /opt/cloudera/parcels/CDH/lib/hive_warehouse_connector/pyspark_hwc-1.0.0.7.1.7.0-551.zip \
--conf spark.sql.extensions="com.hortonworks.spark.sql.rule.Extensions" \
--conf spark.kryo.registrator="com.qubole.spark.hiveacid.util.HiveAcidKyroRegistrator" \
--conf spark.sql.hive.hiveserver2.jdbc.url="jdbc:hive2://devm01.devops.com:2181,devm02.devops.com:2181,devm03.devops.com:2181/defau
--conf spark.sql.hive.hiveserver2.jdbc.url.principal="hive/_HOST@AMDEV.COM" \
--conf spark.security.credentials.hiveserver2.enabled=true \
--conf spark.datasource.hive.warehouse.read.mode=DIRECT_READER_V2
```



```
[leslie@devm01 ~]$ pyspark \
> --jars /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hive-warehouse-connector-assembly-1.0.0.7.1.7.0-551.jar \
> --py-files /opt/cloudera/parcels/CDH/lib/hive_warehouse_connector/pyspark_hwc-1.0.0.7.1.7.0-551.zip \
> --conf spark.sql.extensions="com.hortonworks.spark.sql.rule.Extensions" \
> --conf spark.kryo.registrator="com.qubole.spark.hiveacid.util.HiveAcidKyroRegistrator" \
> --conf spark.sql.hive.hiveserver2.jdbc.url="jdbc:hive2://devm01.devops.com:2181,devm02.devops.com:2181,devm03.devops.com:2181/default;service
DiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2" \
> --conf spark.sql.hive.hiveserver2.jdbc.url.principal="hive/_HOST@AMDEV.COM" \
> --conf spark.security.credentials.hiveserver2.enabled=true \
> --conf spark.datasource.hive.warehouse.read.mode=DIRECT_READER_V2
Python 2.7.5 (default, Oct 14 2020, 14:45:30)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/10/21 16:20:21 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:20:24 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:20:28 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:20:28 WARN ipc.Client: Exception encountered while connecting to the server : org.apache.hadoop.ipc.RemoteException(org.apache.hado
op.ipc.StandbyException): Operation category READ is not supported in state standby. Visit https://s.apache.org/sbnn-error
22/10/21 16:20:36 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.7.7.1.7.0-551
      /_/

Using Python version 2.7.5 (default, Oct 14 2020 14:45:30)
SparkSession available as 'spark'.
>>> 
```

- 使用 HWC 查詢 Internal table 資料

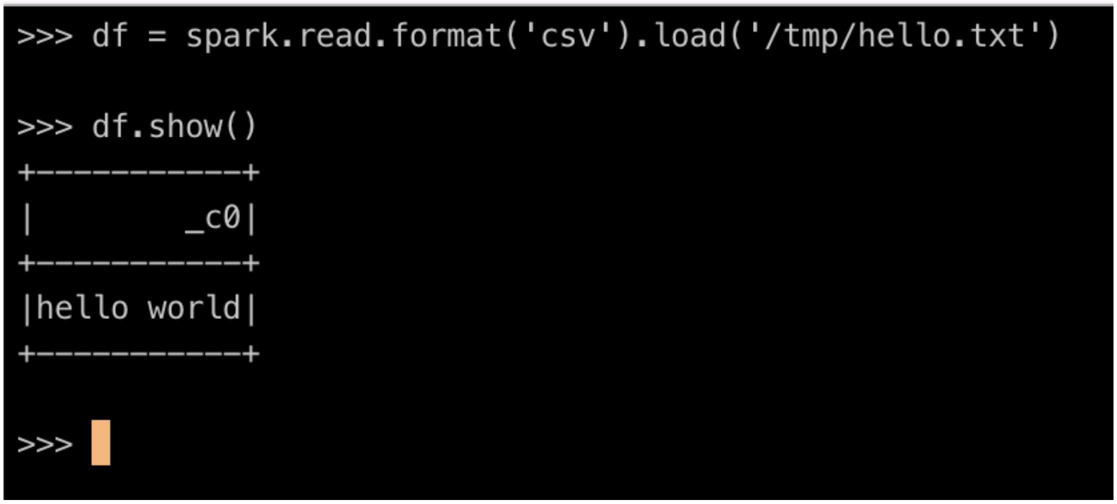
```
from pyspark_llap import HiveWarehouseSession
from pyspark.sql.functions import col
hive = HiveWarehouseSession.session(spark).build()
hive.sql("select * from default.users_internal").show()
```



```
>>> from pyspark_llap import HiveWarehouseSession
>>> from pyspark.sql.functions import col
>>> hive = HiveWarehouseSession.session(spark).build()
>>> hive.sql("select * from default.users_internal").show()
22/10/21 16:24:36 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:24:37 WARN llap.HiveWarehouseSessionImpl:  com.qubole.spark.hiveacid.HiveAcidAutoConvertExtension will be replaced by com.hortonwor
ks.spark.sql.rule.Extensions.Please switch to com.hortonworks.spark.sql.rule.Extensions.
22/10/21 16:24:38 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
Hive Session ID = d7d91867-ee9d-4cda-8d49-63b614702bc2
22/10/21 16:24:40 INFO rule.HWCSwitchRule: using DIRECT_READER_V2 extension for reading
22/10/21 16:24:40 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:24:42 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
22/10/21 16:24:42 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
+------+---+
|  name|age|
+------+---+
| Alice| 16|
|   Bob| 18|
| Cindy| 53|
|   Joe| 28|
| Apple|  5|
|  John| 33|
| Roger| 45|
| Kevin| 17|
|Gordon| 30|
|  Paul| 33|
| Sandy| 39|
|  Alex| 70|
+------+---+
```

- 讀取資料 (HDFS)

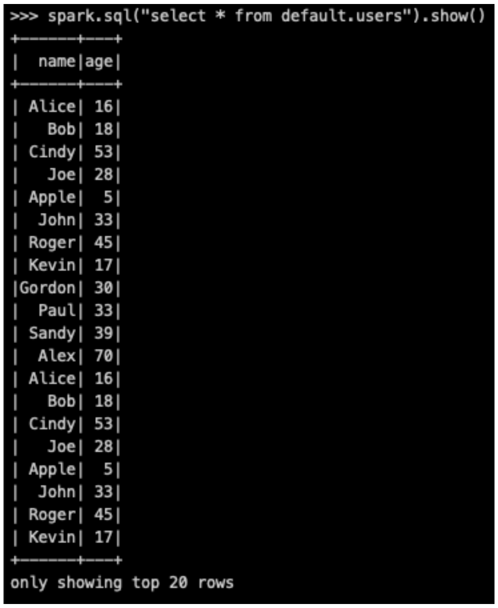
```
df = spark.read.format('csv').load('/tmp/hello.txt')
df.show()
```



```
>>> df = spark.read.format('csv').load('/tmp/hello.txt')

>>> df.show()
+-----------+
|        _c0|
+-----------+
|hello world|
+-----------+

>>> 
```

- 讀取資料 (External table)

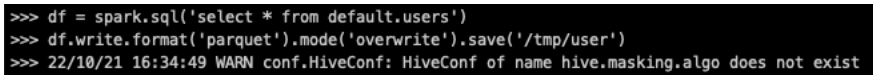
```
spark.sql("select * from default.users").show()
```



```
>>> spark.sql("select * from default.users").show()
+------+---+
|  name|age|
+------+---+
| Alice| 16|
|   Bob| 18|
| Cindy| 53|
|   Joe| 28|
| Apple|  5|
|  John| 33|
| Roger| 45|
| Kevin| 17|
|Gordon| 30|
|  Paul| 33|
| Sandy| 39|
|  Alex| 70|
| Alice| 16|
|   Bob| 18|
| Cindy| 53|
|   Joe| 28|
| Apple|  5|
|  John| 33|
| Roger| 45|
| Kevin| 17|
+------+---+
only showing top 20 rows
```

- 寫入資料至 HDFS

```
df = spark.sql('select * from default.users')
df.write.format('parquet').mode('overwrite').save('/tmp/user')
```



```
>>> df = spark.sql('select * from default.users')
>>> df.write.format('parquet').mode('overwrite').save('/tmp/user')
>>> 22/10/21 16:34:49 WARN conf.HiveConf: HiveConf of name hive.masking.algo does not exist
```

▼ Python 3

```
python3 --version
```

▼ Java 8

```
java -version

javac -version
```

Executor List

1. bin/application.properties
2. bin/jdbc-kerberos-0.0.1-SNAPSHOT.jar
3. README.md

Reference

1. jdbc-doc/Cloudera-JDBC-Driver-for-Apache-Hive-Install-Guide.pdf
2. jdbc-doc/Cloudera-JDBC-Driver-for-Impala-Install-Guide.pdf
3. source_code (jdbc-kerberos)