模块五 Hive数据操作



《Hive数据仓库应用教程》

目 录 / CONTENTS

1 任务5-1 数据导入

2 任务5-2 数据导出

3 任务5-3 本章小结

4 任务5-4 本章习题

任务 5-2 数据导出

5.2.1 任务描述

在Hive的日常使用中，经常需要将Hive表中的数据导出来，虽然Hive提供了多种导出方式，但是面对不同的数据量、不同的需求，如果随意使用某种导出方式，可能会导致导出时间过长，导出的结果不满足需求，甚至造成集群资源不必要的浪费。

因此本小节主要对Hive支持的几种导出方式的使用进行说明，并给出每种导出方式的使用场景，便于指导学生能够选取最佳的导出方式。







5.2.1 任务准备知识

从Hive数据仓库中导出数据到本地，因为Hive的存储依赖于HDFS分布式文件系统，所以从Hive中导出数据其实就是从HDFS中导出数据。

由于此处语法格式中的关键字和相关代表词的含义本章第一小节数据的导入部分已经做了详细的说明，故此不在此赘述，请读者注意。

5.2.3 任务实施

Hive 数据导出方式，和导入方式一样，针对不同场景有很多方式，主要方式有

- Insert 导出、
- Hadoop 命令导出、
- Hive Shell 命令导出、
- Export 命令导出

等四种方式。

一、 Insert 语句导出

这种加载数据的方式一般是将查询到的数据加载到Hadoop 分布式文件系统中。

语法格式：

```
INSERT  OVERWRITE  [LOCAL] DIRECTORY  'filepath'
SELECT  select_statement1  FROM  from_statement;
```

5.2.3 任务实施

例题 14：将表 person 中的数据导出到本地文件系统/usr/datas/out1 目录中。

操作过程：

step01：　　查看本地文件系统的目录结构

查看本地文件系统的目录结构，以方便确定待导出数据的存放路径。具体操作命令如下：

```
[root@node2 ~]#    cd /usr/datas/
[root@node2 datas]#ll
```

其中，本地文件系统的目录结构，其中并没有文件目录 out1，此目录为本题目待导出数据文件的存储目录。

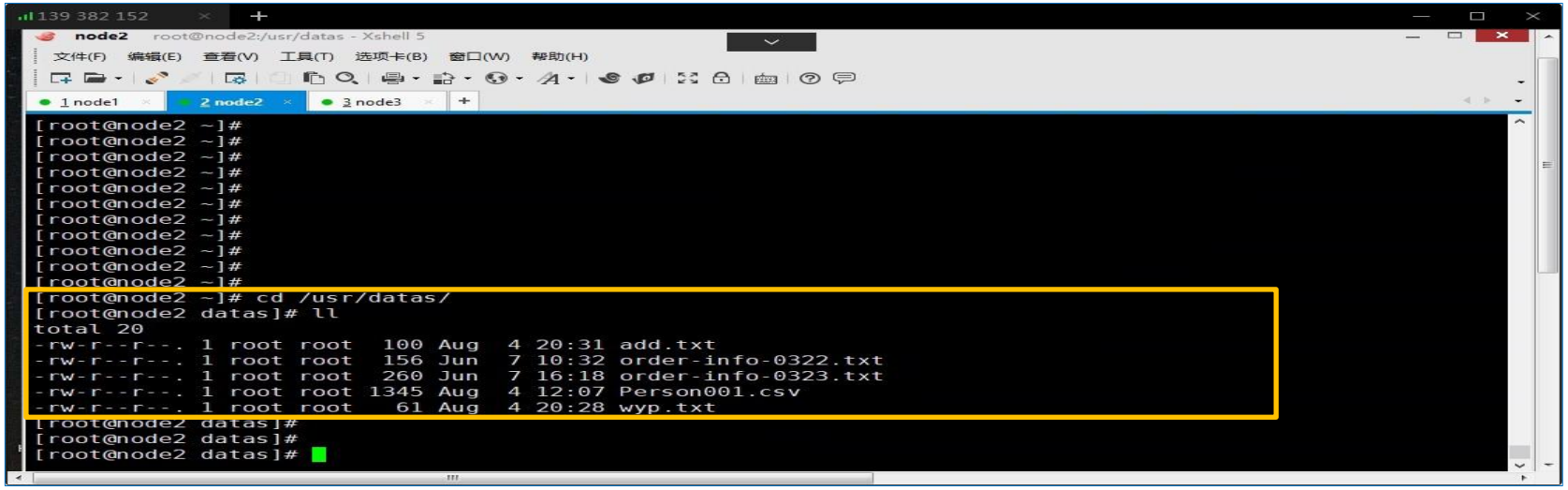


图 5-49 查看 HDFS 文件系统的目录

5.2.3 任务实施

例题 14：将表 person 中的数据导出到本地文件系统/usr/datas/out1 目录中。

操作过程：

step02：查看 person 表中的数据。查看 person 表中的数据，由图 5-50 和图 5-51 所示，person 表中共有 80 条记录，此处仅显示 11 条记录。
具体操作命令如下：
hive>SELECT * FROM person;

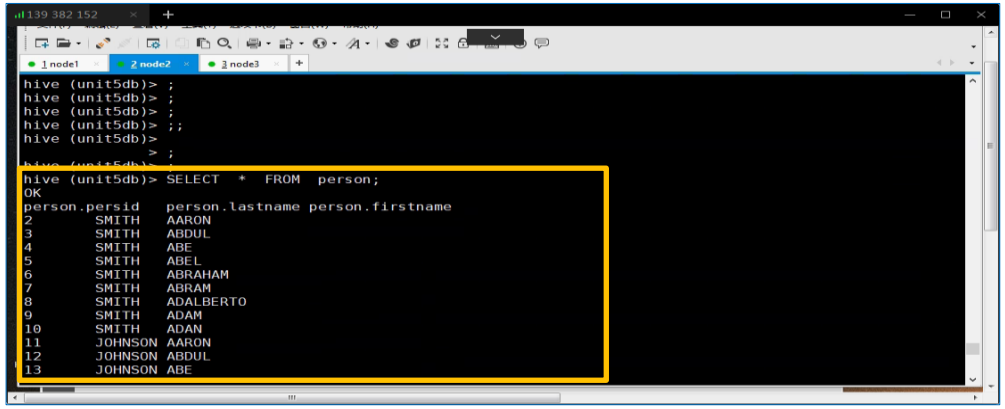


图 5-50 person 表中的数据一

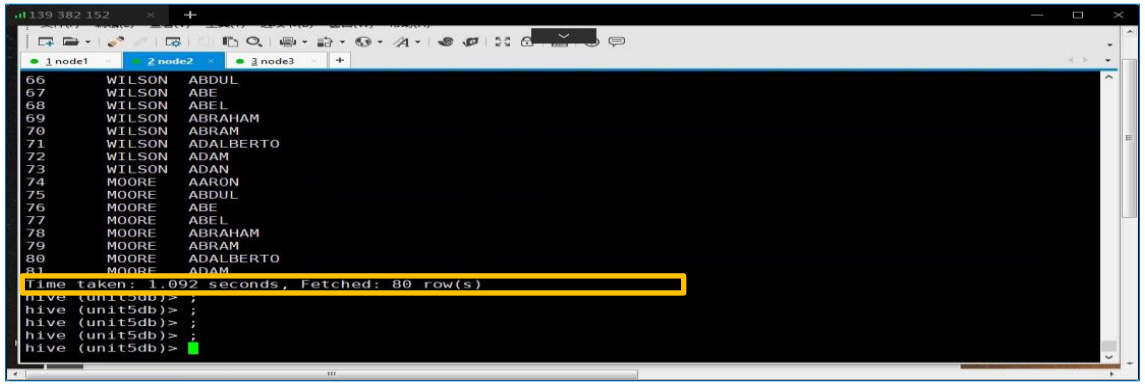


图 5-51 person 表中的数据一

step03：　　查询 person 表中的数据，并写入文件系统

然后，按照查询要求查询 person 表中的数据，并且写入 HDFS 文件系统 /tmp/person-hdfs 中。

具体操作命令如下：

```
hive>INSERT OVERWRITELOCAL DIRECTORY '/usr/datas/out1'
SELECT persid, firstname, lastname FROM person;
```

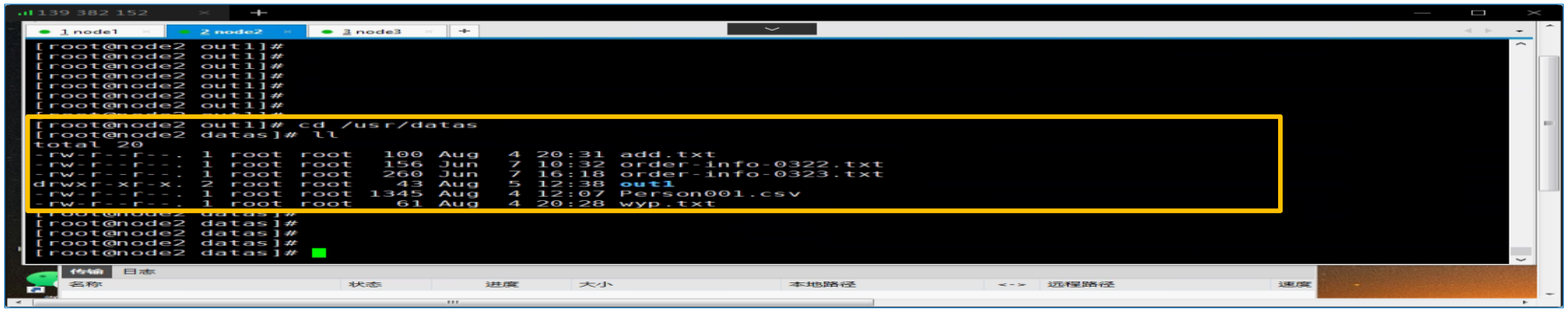


图 5-52 再次查看/usr/datas/的目录

step04：　　验证/usr/datas/目录中的数据

再次查看本地文件系统中的目录结构/usr/datas/的目录，

结果如图 5-52 所示，

由结果可知/usr/datas/out1 已经存在，即待导出数据文件的存储目录/out1 是自动生成的。

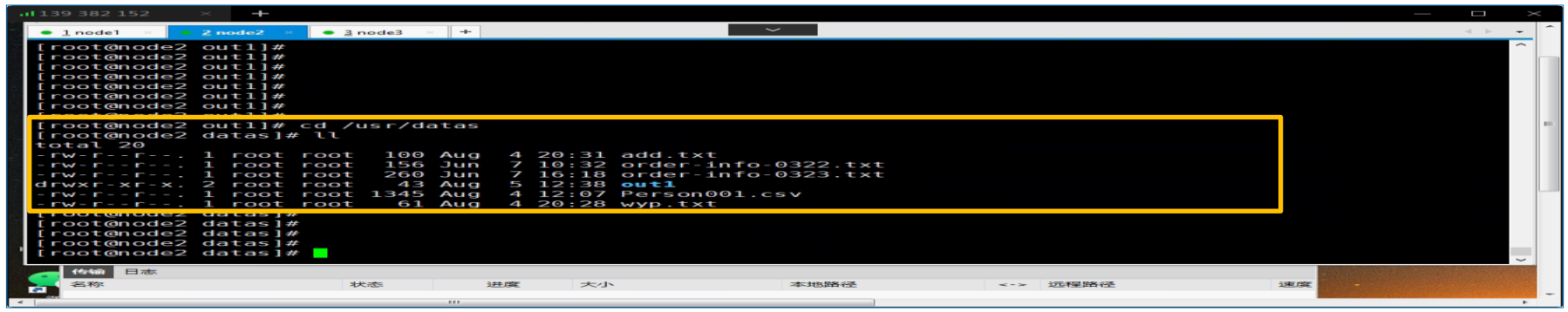


图 5-52 再次查看/usr/datas/的目录

5.2.3 任务实施

验证数据是否导出到存储目录，具体操作命令如下：

```
[root@node2 datas]#  cd  /usr/datas/out1
[root@node2 out1]#  ll
[root@node2 out1]#  cat  000000_0
```

由结果可知，如图 5-53 和图 5-54 所示，整个/out1 中共有 80 条目录，数据导出成功。

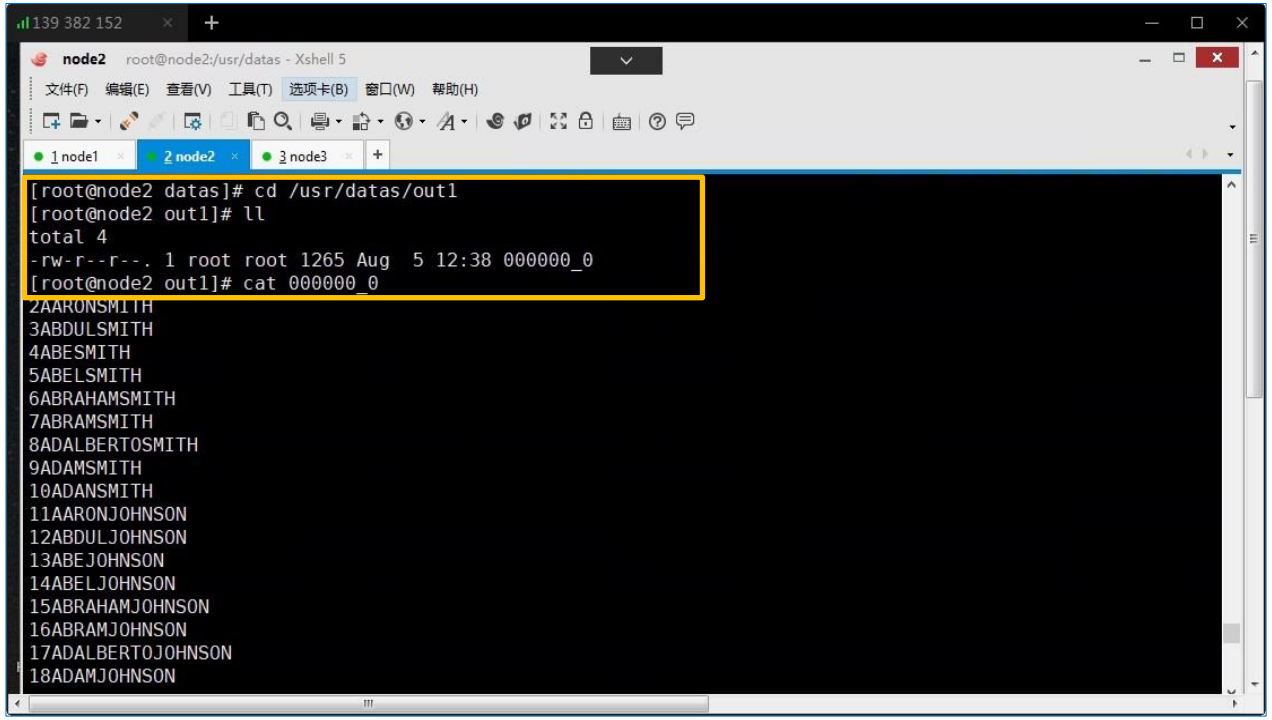


图 5-53 查看/usr/datas/out1 中的数据一

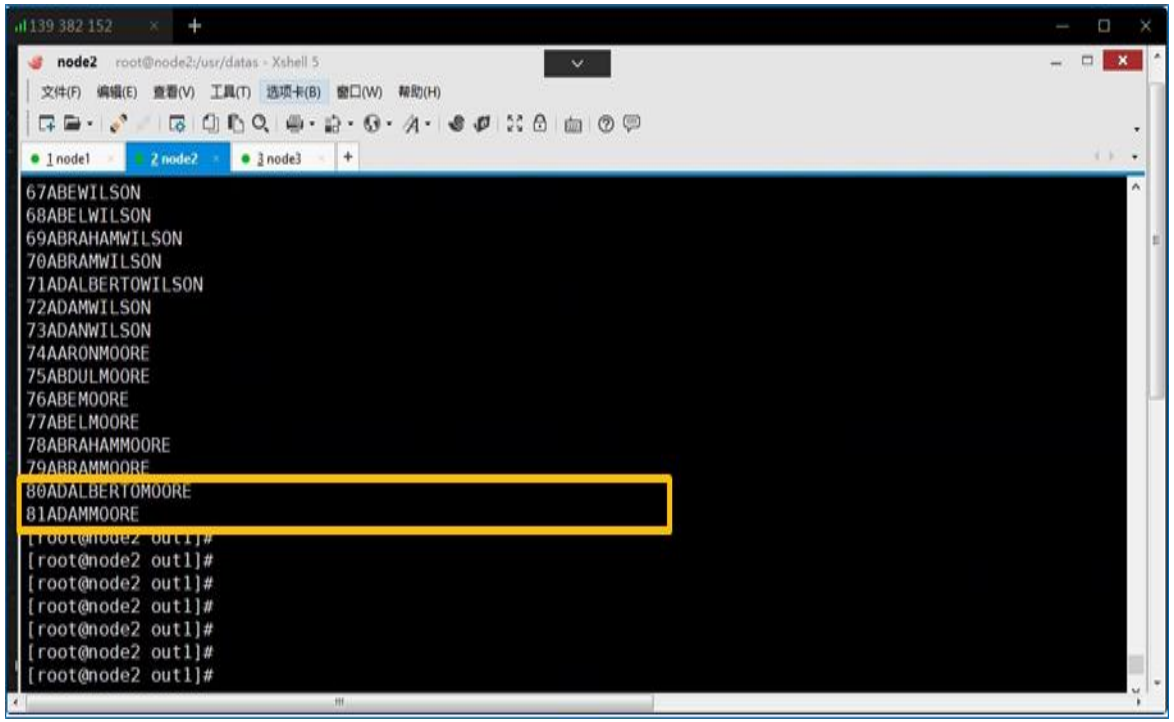


图 5-54 查看/usr/datas/out1 中的数

5.2.3 任务实施

例题 15：将表 person 中的数据按照格式化标准导出到本地文件系统/usr/datas/out2目录中。

操作过程：

step01：　　查看本地文件系统的目录结构。查看本地文件系统的目录结构，以方便确定待导出数据的存放路径。具体操作命令如下：

[root@node2 ~]#　cd /usr/datas/ [root@node2 datas]#　ll

其中，本地文件系统的目录结构，如图 5-55 所示。共有 20 个文件和目录，但是，其中并没有文件目录 out2，此目录为本题目待导出数据文件的存储目录。

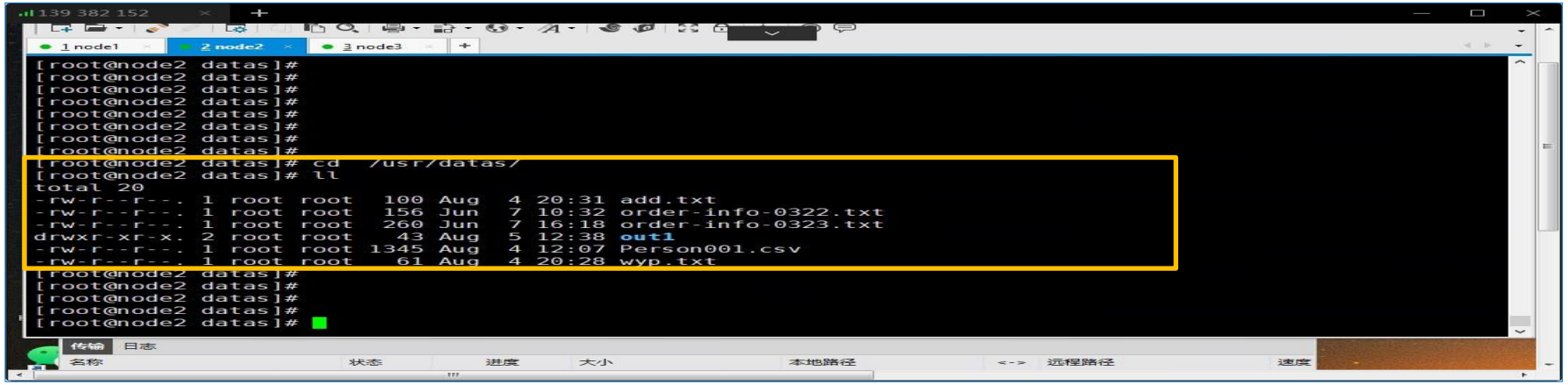


图 5-55 查看 HDFS 文件系统的目录

5.2.3 任务实施

step02：查看 person 表中的数据。查看 person 表中的数据，由图 5-48 和图 5-49 所示，person 表中共有 80 条记录。具体操作命令如下：hive> SELECT * FROM person;

step03：查询 person 表中的数据，并写入文件系统。然后，按照查询要求查询 person 表中的数据，即待导出数据文件的存储目录/out2 是自动生成的。具体操作命令如下：

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/usr/datas/out2' ROW FORMAT
    DELIMITED FIELDS TERMINATED by ',' SELECT persid, firstname, lastname
    FROM person;
```

step04：验证/usr/datas/目录中的数据。再次查看本地文件系统中的目录结构/usr/datas/的目录，结果如图 5-56 所示，由结果可知/usr/datas/out2 已经存在，即待导出数据文件的存储目录/out2 是自动生成的。具体操作命令如下：

```
[root@node2 ~]# cd /usr/datas/ [root@node2 datas]#
```

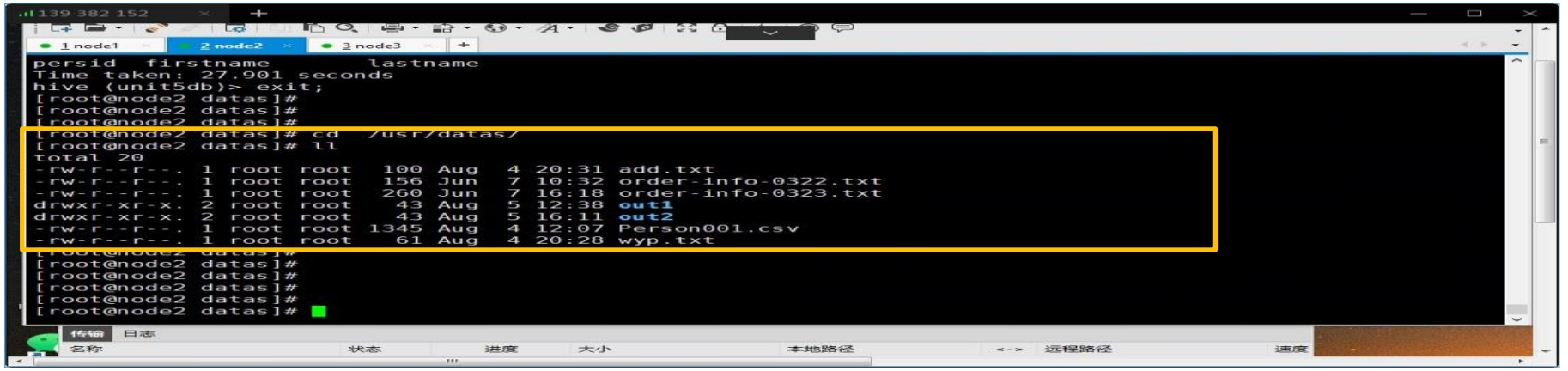


图 5-56 再次查看/usr/datas/的目录

5.2.3 任务实施

验证数据是否导出到存储目录，具体操作命令如下：
[root@node2 datas]# cd /usr/datas/out2 root@node2 out2]# ll [[root@node2 out2]# cat 000000_0
由结果可知，如图5-57和图5-58所示，整个/out2中共有80条目录，数据导出成功。

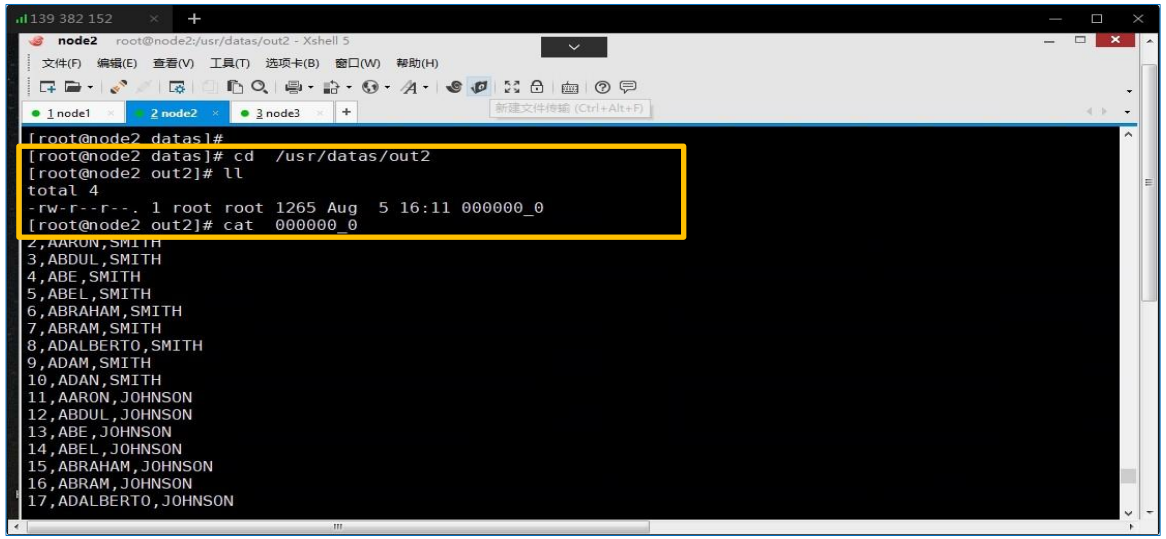


图 5-57 查看/usr/datas/out2 中的数据一

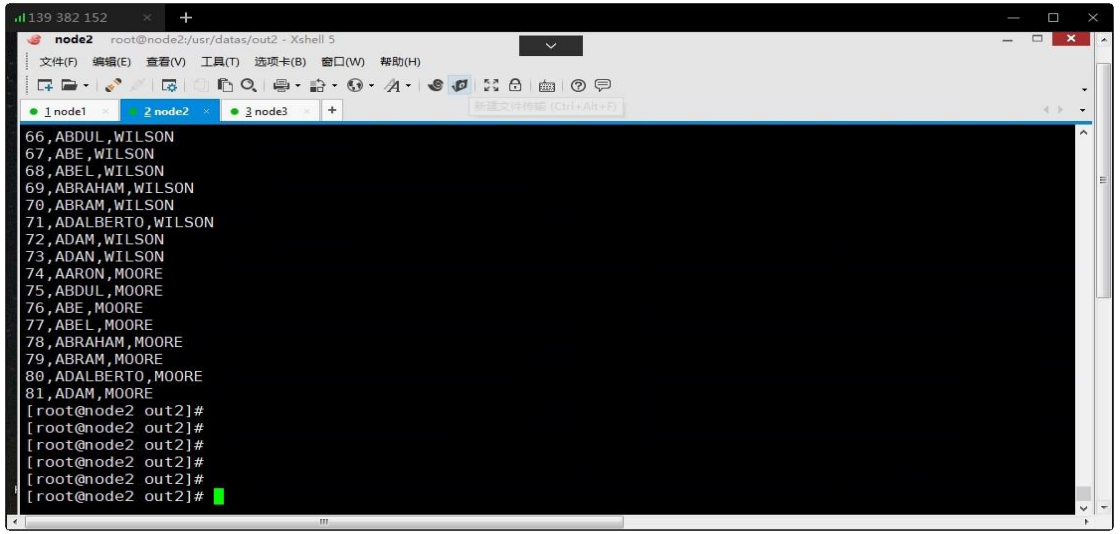


图 5-58 查看/usr/datas/out2 中的数据二

5.2.3 任务实施

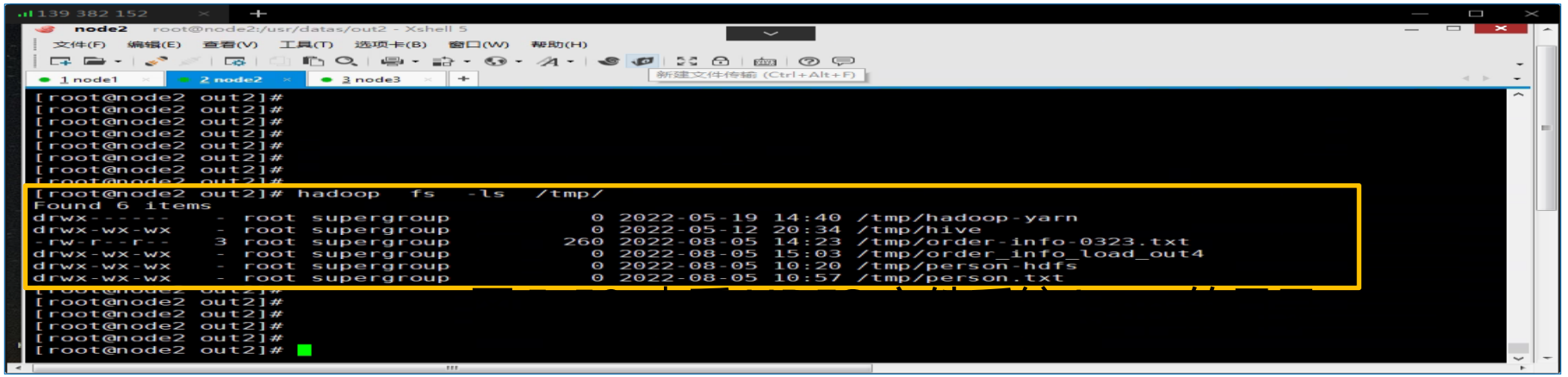
例题 16：将表 person 中的数据导出到 HDFS 的文件系统/tmp/out3 目录中。

操作过程：

step01：查看 HDFS 的文件系统/tmp 的目录结构。查看 HDFS 的文件系统/tmp 的目录结构，以方便确定待导出数据的存放路径。具体操作命令如下：

[root@node2 ~]#　hadoop fs　　-ls　　/tmp/

其中，HDFS 的文件系统/tmp 的目录结构，如图 5-59 所示。共有 6 个目录，但是，其中并没有文件目录 out3，此目录为本题目待导出数据文件的存储目录。



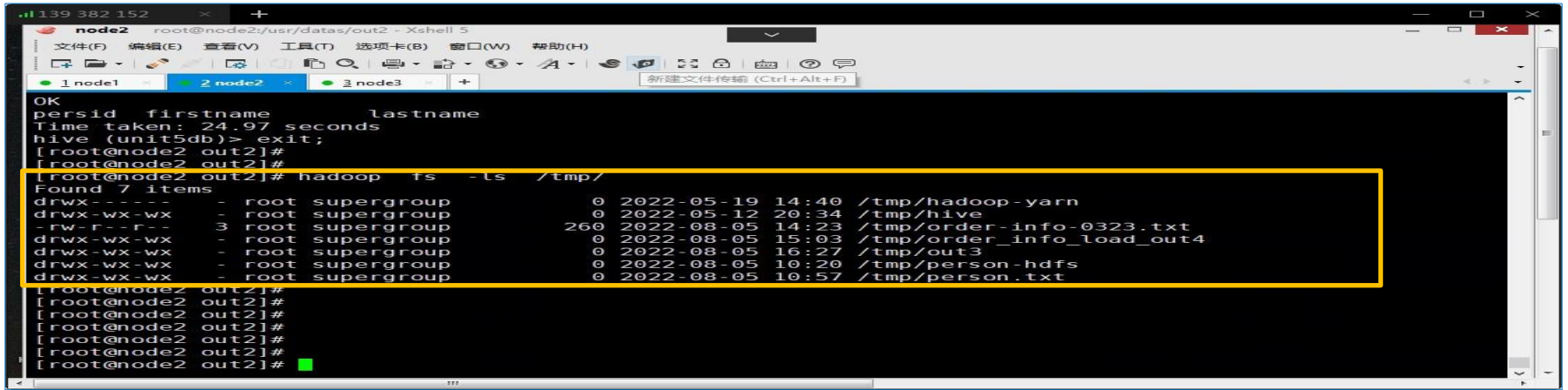
5.2.3 任务实施

step02：　查看 person 表中的数据。查看 person 表中的数据，由图 5-48 和图 5-49 所示，person 表中共有 80 条记录。具体操作命令如下：hive>　SELECT *　FROM　person;

step03：　查询 person 表中的数据，并写入文件系统。然后，按照查询要求查询 person 表中的数据，并且写入 HDFS 文件系统/tmp/person-hdfs 中。具体操作命令如下：

```
hive>　INSERT　OVERWRITE　DIRECTORY　'/tmp/out3'
　　ROW　FORMAT　DELIMITED　FIELDS TERMINATED　by ','
　　SELECT　persid,　firstname,　lastname　FROM　person;
```

step04：　验证/tmp/目录中的数据。再次查看本地文件系统中的目录结构/tmp/的目录，结果如图 5-60 所示，具体操作命令如下：[root@node2 ~]#　hadoop　fs　-ls /tmp/



5.2.3 任务实施

由结果可知/tmp/out3 已经存在，即待导出数据文件的存储目录/out3 是自动生成的，且/tmp/out3 存在文件，如图 5-61 所示。具体操作命令如下：

[root@node2 ~]# hadoop fs -ls /tmp/out3

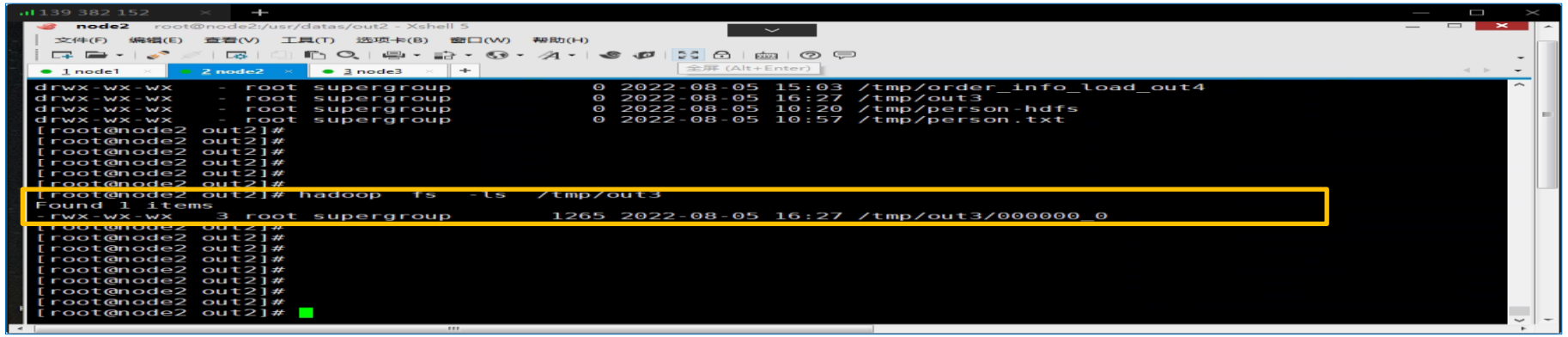


图 5-61 查看/tmp/out3 的目录

验证数据是否导出到存储目录，具体操作命令如下：[root@node2 ~]# cd /usr/datas/。进入本地文件系统/usr/datas/

[root@node2 datas]# Hadoop fs -get /tmp/out3/

下载/tmp/out3/中的内容到本地文件系统/usr/datas/中

[root@node2 datas]# ls [root@node2 datas]# cd out3 [root@node2 out3]# ls

[root@node2 out3]# cat 000000_0

5.2.3 任务实施

由结果可知/tmp/out3 已经存在，即待导出数据文件的存储目录/out3 是自动生成的，且/tmp/out3 存在文件，如图 5-61 所示。具体操作命令如下：

[root@node2 ~]# hdfs dfs –ls /tmp/out3

验证数据是否导出到存储目录，具体操作命令如下：

```
[root@node2 ~]#cd /usr/datas/
进入本地文件系统 /usr/datas/

[root@node2 datas]#hdfs dfs  -get  /tmp/out3/

下载/tmp/out3/中的内容到本地文件系统/usr/datas/中
[root@node2 datas]#ls
[root@node2 datas]#cd  out3
[root@node2 out3]#ls
[root@node2 out3]#cat  000000_0
```

5.2.3 任务实施

由结果可知，如图 5-62 和图 5-63 所示，整个/out3 中共有 80 条目录，数据导出成功。

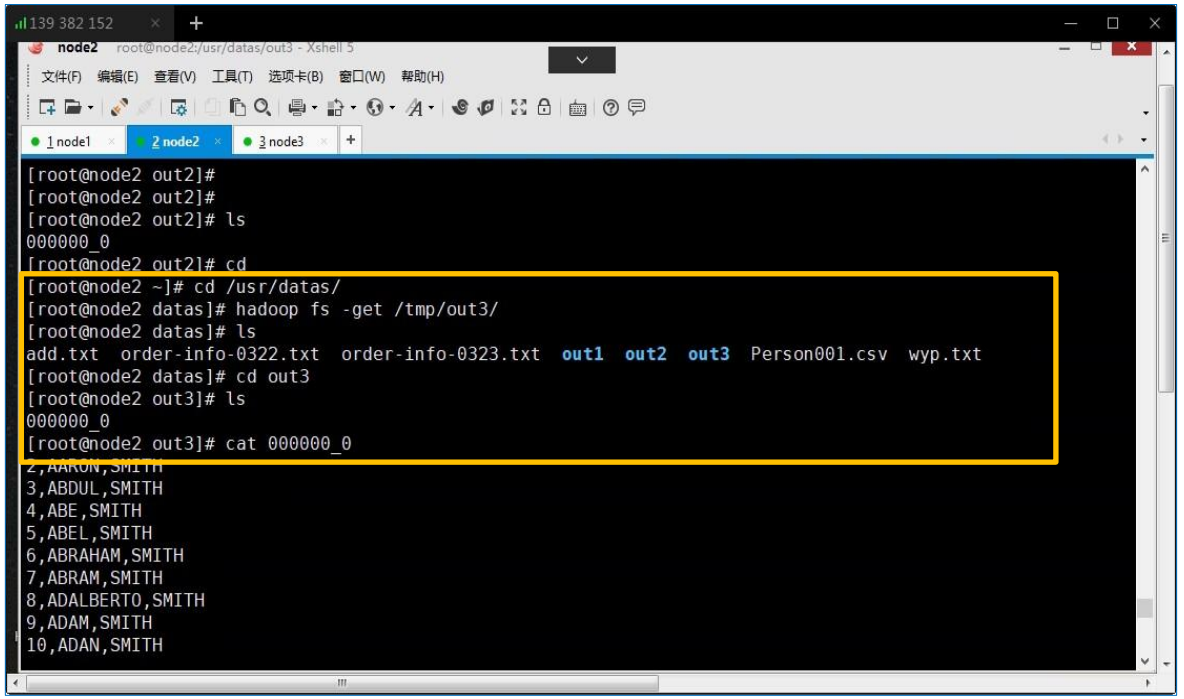


图 5-62 查看/usr/datas/out3 中的数据一

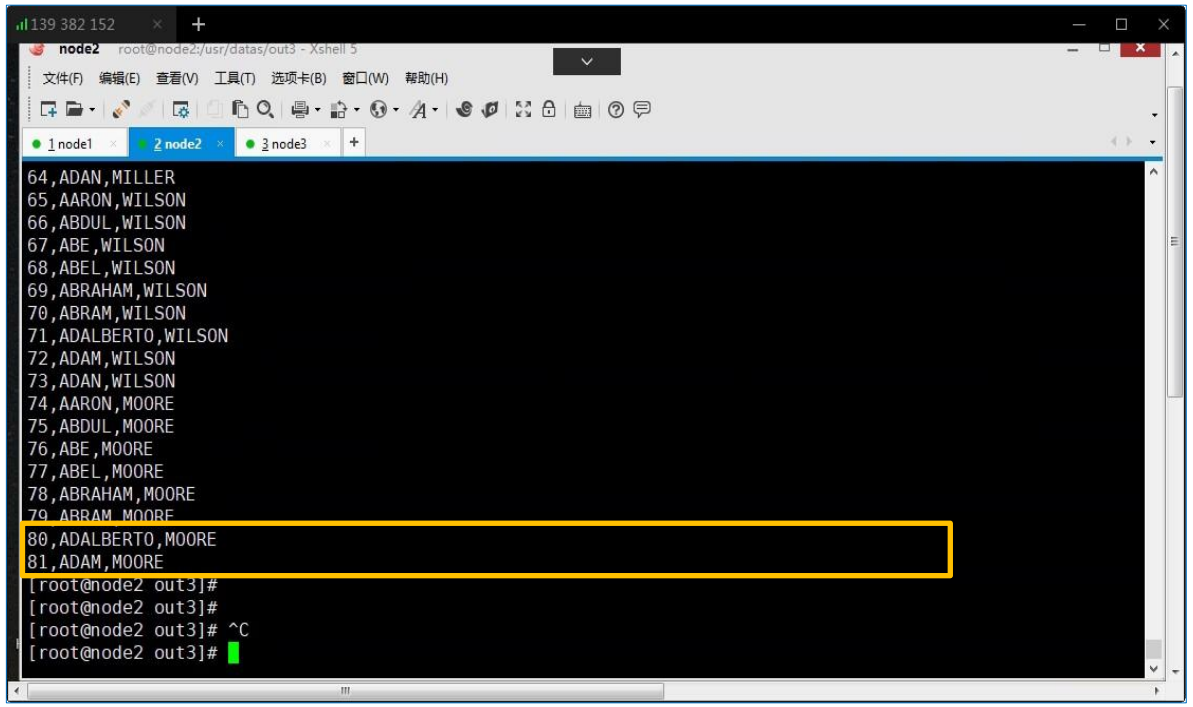
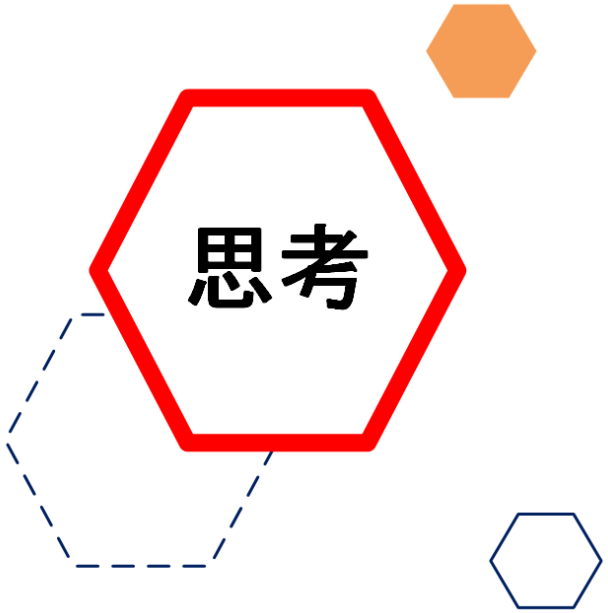


图 5-63 查看/usr/datas/out3 中的数据二



使用场景：

1）由于这种方式导出数据时会启动MapReduce任务来完成导出，所以这种方式不适合导出数据量极少的情况，因为数据量很少时也会启MapReduce任务消耗集群资源，且导出速度比没有启用MapReduce导出方式的速度要慢；

2）适用于对导出数据格式有要求的场景，如：指定数据文件类型、字段分割符；

3）这种方式导出的路径只能指定到文件夹，不支持导出到指定的文件中，且在文件夹下可能产生多个文件，并非仅一个文件，所以此方式适用于将数据导出到本地或HDFS目录下，且对存储数据的文件名和文件数量无要求。

二、Hadoop命令导出

Hadoop命令在使用时，一般用于将HDFS文件系统中的数据导出到本地文件系统中。

语法格式：

hdfs dfs -get [文件目录] [本地目录]

5.2.3 任务实施

例题 17：将HDFS 上的文件/tmp/out3 中的数据导出到本地/usr/datas/hadoop 目录中。

操作过程：

step01：查看 HDFS 的文件系统/tmp 的目录结构。

查看 HDFS 的文件系统/tmp 的目录结构，以便于确定例题 16 中得出的/out3目录仍然存在。

具体操作命令如下：

```
hdfs dfs -ls /tmp/
```

其中，HDFS 的文件系统/tmp 的目录结构，如图 5-64 所示。共有 6 个目录，其中存在文件目录out3，此目录为本题目待下载的数据的存储目录。

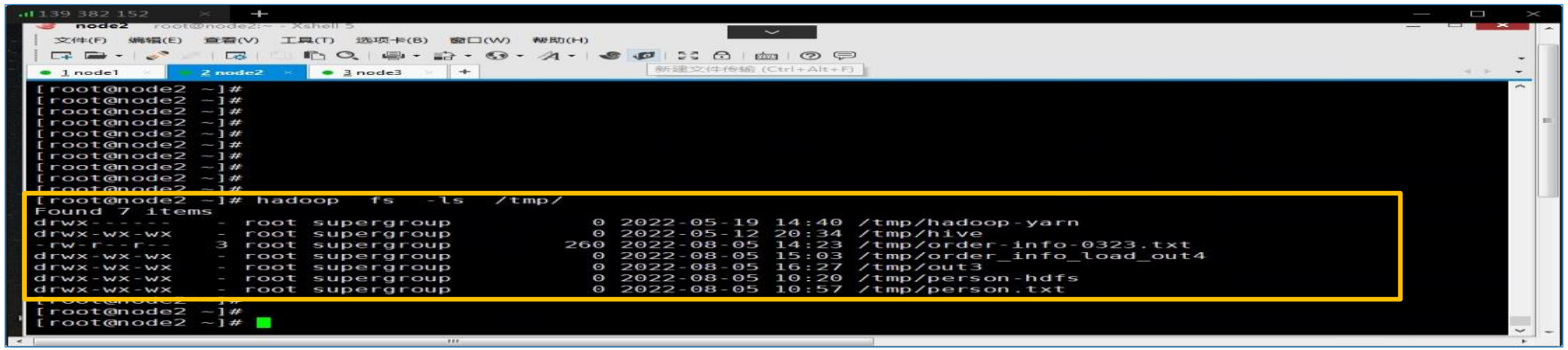


图 5-64 查看 HDFS 文件系统/tmp 的目录

5.2.3 任务实施

查看 HDFS 的文件系统/tmp/out3 的目录结构，以便于确定例题 16 中得出的数据仍然存在。

具体操作命令如下：

[root@node2 ~]#　hdfs dfs -ls /tmp/out3

其中，HDFS 的文件系统/tmp/out3 的目录结构，

如图 5-65 所示。共有 1 个目录，所需数据仍然存在。

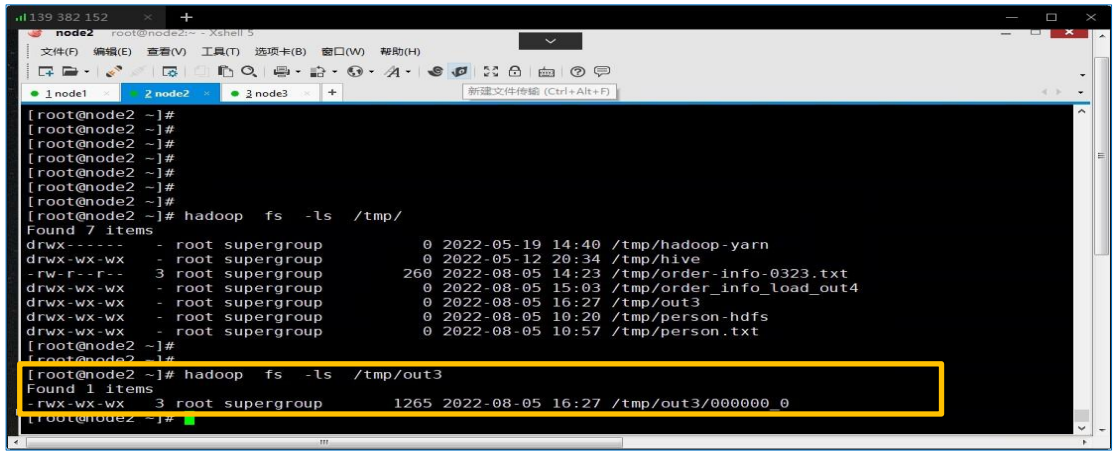


图 5-65 查看 HDFS 文件系统/tmp/out3 的目录

5.2.3 任务实施

step02：下载/tmp/out3 目录中的数据到本地。
再次查看本地文件系统中的目录结/usr/datas 的目录，
结果如图 5-66 所示，
具体操作命令如下：

```
[root@node2 ~]#  cd  /usr/datas
[root@node2 datas]#  ll
```

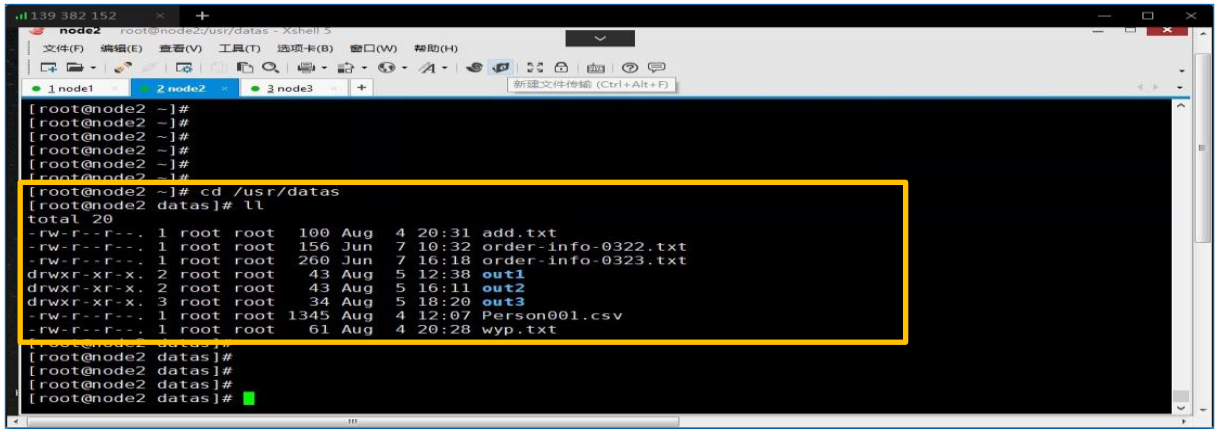


图 5-66 查看/usr/datas 的目录

5.2.3 任务实施

其中，目录/hadoop 不存在，
则创建目录/hadoop，
使其本地路径为/usr/datas/hadoop，
如图 5-67 所示，
具体操作命令如下：

[root@node2 datas]#mkdir /usr/datas/hadoop
[root@node2 datas]#ll

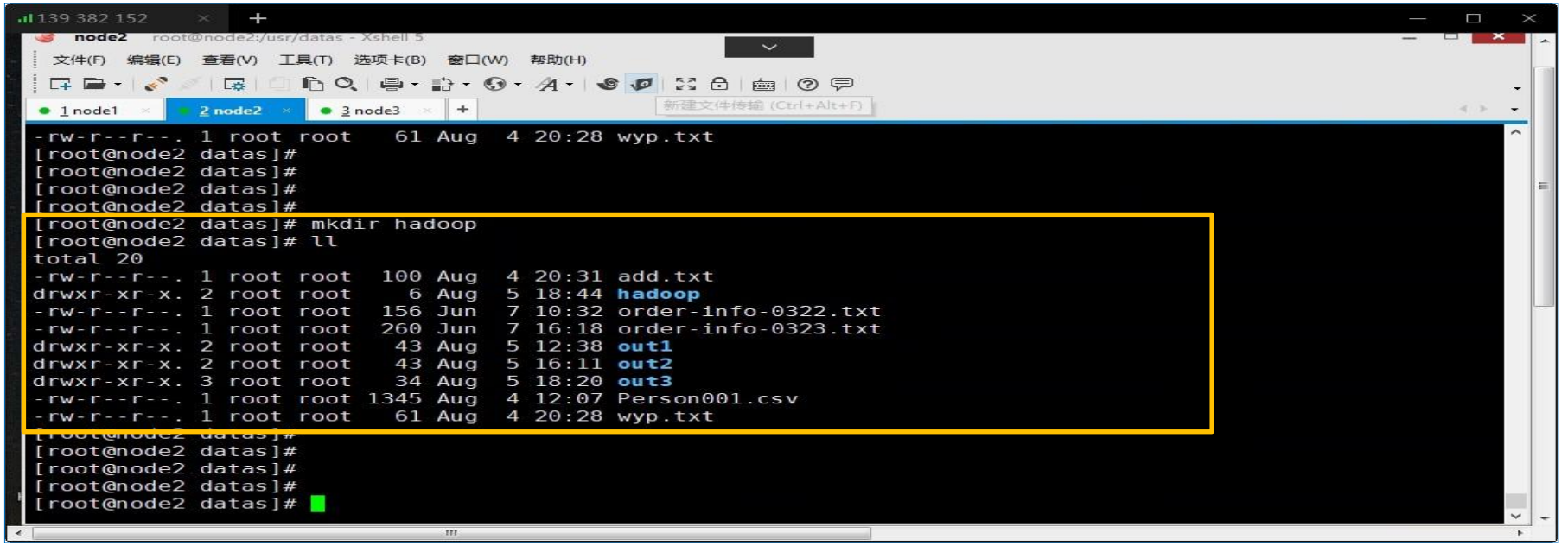


图 5-67 再次查看/usr/datas 的目录

然后，利用 hadoop 命令下载 HDFS 文件系统/tmp/out3 中的数据到本地文件系统/usr/datas/Hadoop 中，

具体操作命令如下：

[root@node2 datas]#hdfs dfs -get /tmp/out3 /usr/datas/hadoop

step03：验证本地文件系统/usr/datas/hadoop 的数据

查看本地文件系统中的目录结构/usr/datas/hadoop 的目录，

结果如图 5-68 所示，由结果可知，/out3 目录中的数据被下载到本地，

具体操作命令如下：

[root@node2 datas]#cd /usr/datas/hadoop

[root@node2 hadoop]#ll

5.2.3 任务实施

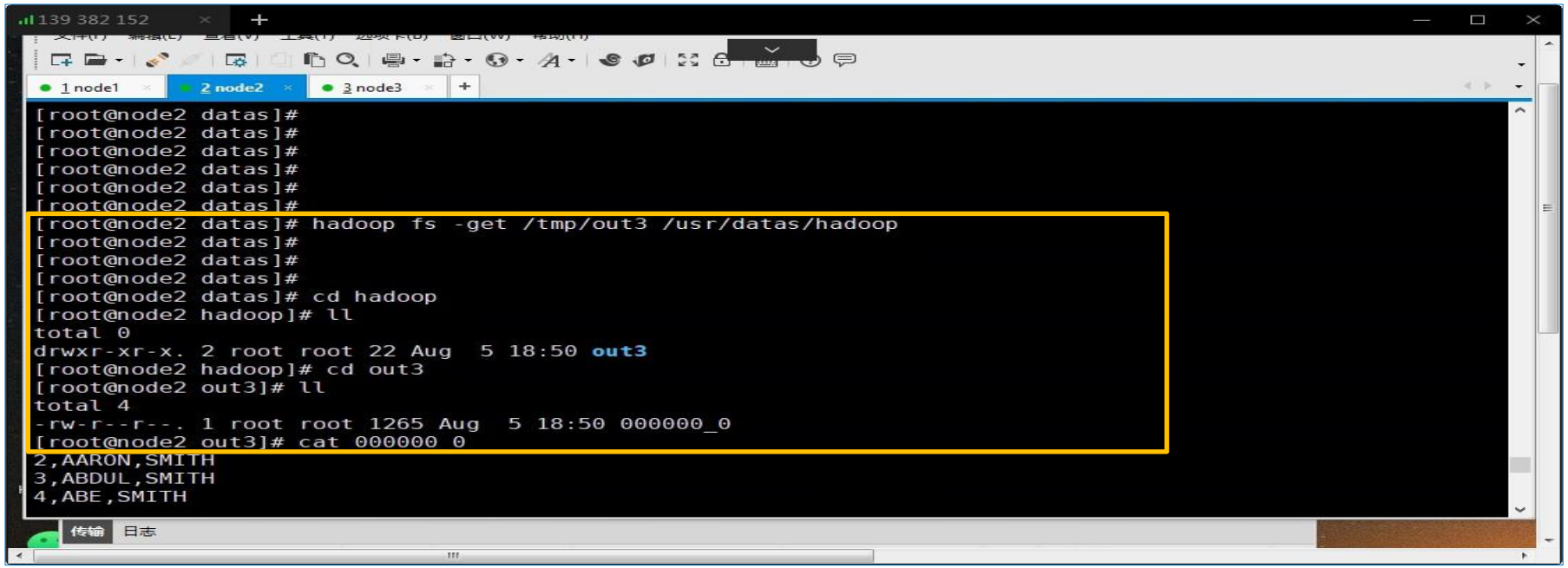


图 5-68 查看/usr/datas/hadoop 的目录

5.2.3 任务实施

查看/usr/datas/hadoop 目录中的数据，共有 80 条数据，

如图 5-69 和图 5-70所示，具体操作命令如下：

[root@node2 hadoop]#cd out3

[root@node2 out3]#ll

[root@node2 out3]#cat 000000_0

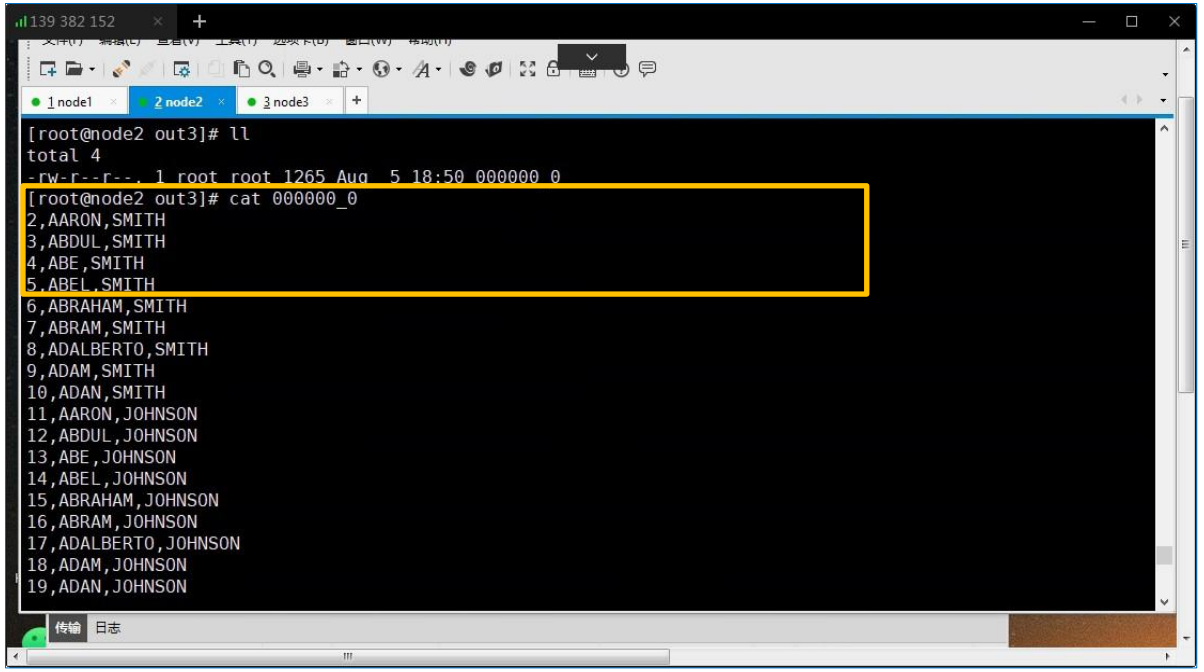


图 5-69 查看/usr/datas/hadoop 的目录一

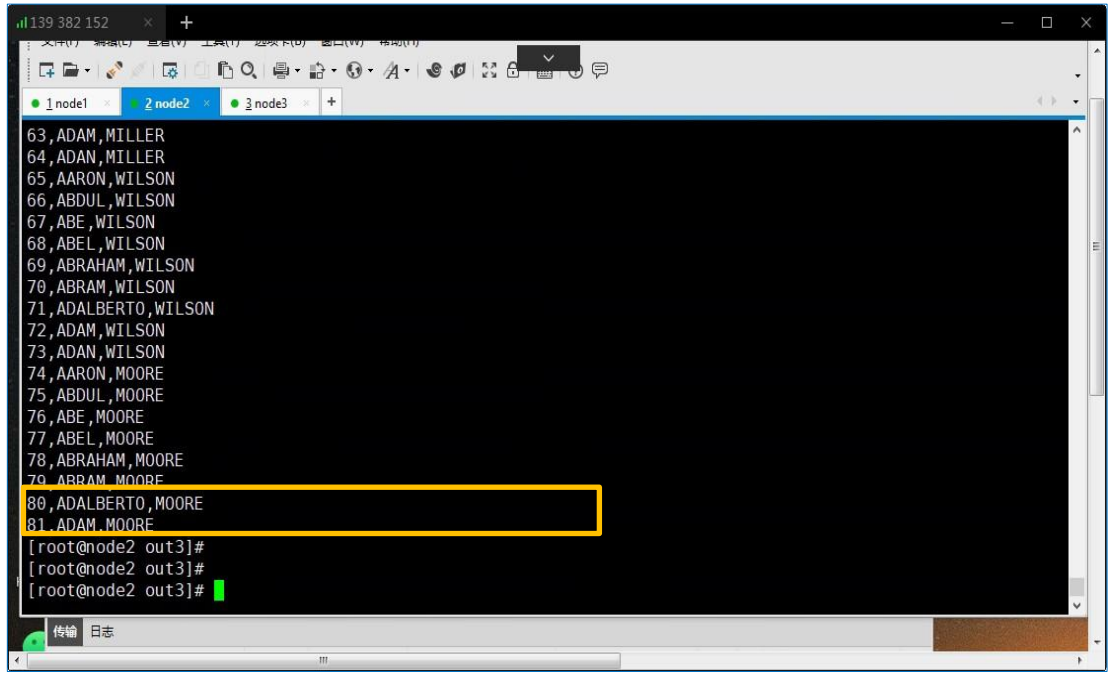


图 5-70 查看/usr/datas/hadoop 的目录二

三、Hive Shell命令导出

还有一种方式是直接使用hive命令，加入参数-e/-f来导出到本地文件。

语法格式：

HIVE -e "sql语句" > filepath

HIVE -f sql文件 > filepath

5.2.3 任务实施

例题 18：将表 order_info_load 中的数据导出到本地文件目录/usr/datas/hiveshell中。

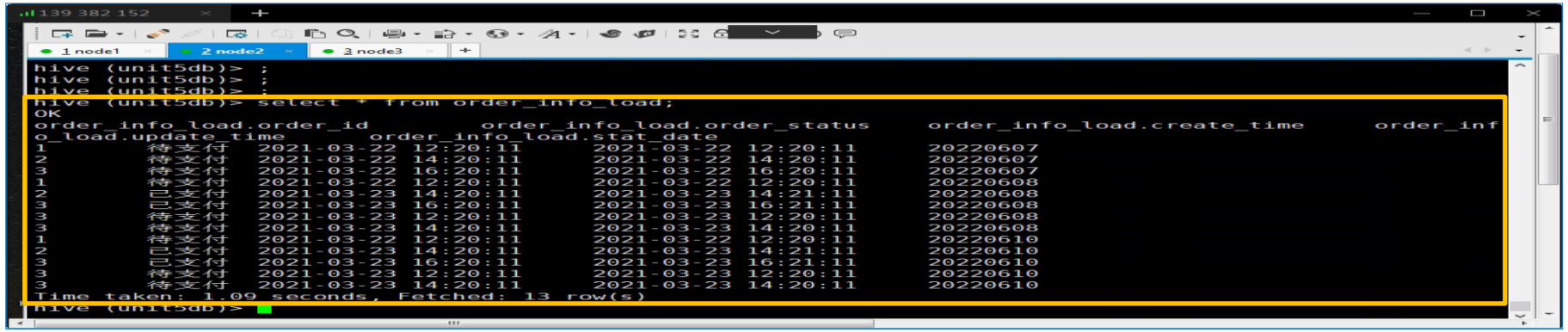
操作过程：

step01：查看表 order_info_load 的数据。查看表 order_info_load 中的数据。

具体操作命令如下：

```
SELECT * FROM order_info_load ;
```

其中，表 order_info_load 中的共有 13 条数据，如图 5-71 所示。



step02：创建本地路径/usr/datas/hiveshell。
查看本地文件系统中的目录结构/usr/datas 的目录，
具体操作命令如下：

```
cd /usr/datas
ll
```

其中，目录/hiveshell 不存在，

则创建目录/hiveshell，

使其本地路径为/usr/datas/hiveshell，

具体操作命令如下：

```
mkdir hiveshell
ll
```

5.2.3 任务实施

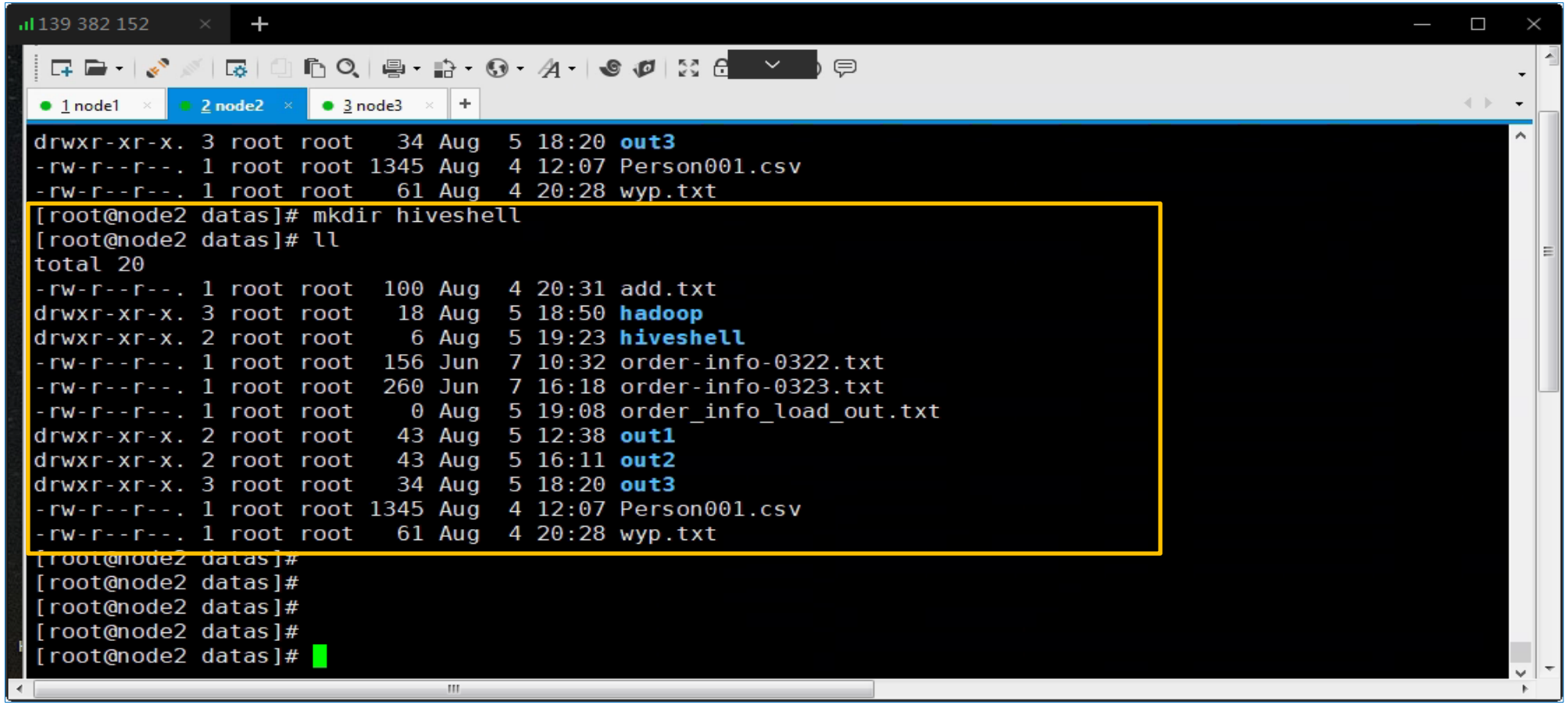


图 5-72 查看/usr/datas 的目录

5.2.3 任务实施

step03：　　验证本地路径/usr/datas/hiveshell 中的数据。利用 Hive Shell 中-e 命令，导出查询数据到本地/usr/datas/hiveshell 目录中，具体操作命令如下：

```
hive　　-e　'select * from unit5db.order_info_load; '>
/usr/datas/hiveshell/order_info_load_out.txt
```

然后，进入本地目录/usr/datas/hiveshell，

查看/order_info_load_out.txt是否存在，

且查看/order_info_load_out.txt 文件中的数据，如图 5-73 所示，

具体操作命令如下：

```
[root@node2 datas]#  cd  hiveshell
[root@node2 hiveshell]#  ll
[root@node2 hiveshell]#  cat  order_info_load_out.txt
```

5.2.3 任务实施

step03：　　验证本地路径/usr/datas/hiveshell 中的数据。利用 Hive Shell 中-e 命令，导出查询数据到本地/usr/datas/hiveshell 目录中，具体操作命令如下：

```
hive　　-e 'select * from unit5db.order_info_load; '>
/usr/datas/hiveshell/order_info_load_out.txt
```

然后，进入本地目录/usr/datas/hiveshell，查看/order_info_load_out.txt是否存在，且查看/order_info_load_out.txt 文件中的数据，如图 5-73 所示，具体操作命令如下：

```
[root@node2 datas]# cd hiveshell [root@node2 hiveshell]# ll
[root@node2 hiveshell]# cat order_info_load_out.txt
```

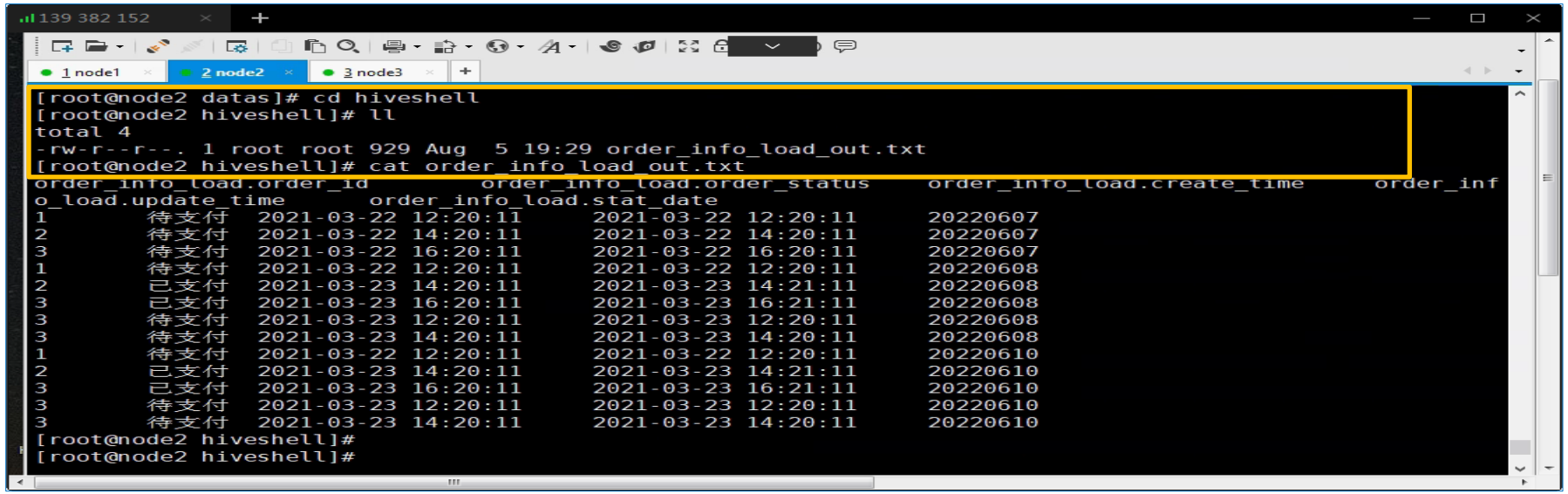
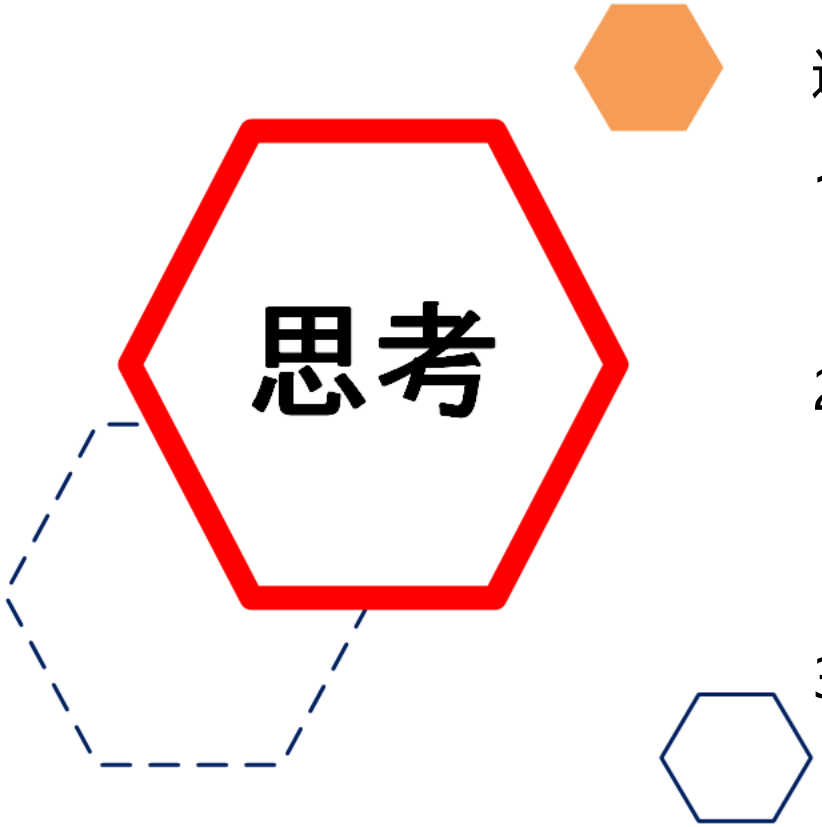


图 5-73 查看/usr/datas/hiveshell 中的数据



适用场景：

1) 这种方式导出数据时不会启动 MapReduce 任务，适用于数据量较少的情况；

2) 此方式支持指定导出的文件名，导出的结果只能存放到本地，且只生成一个文件，所以适用于将hive 数据导出到本地的某个文件中的场景。

3) 不支持指定导出文件的分割符。

四、 Export 语句导出

如前所述，这种导出数据的方式一般先用 Export 将数据导出后，再将数据使用Import 导入，利用Export 和Import 主要用于两个Hadoop 平台集群之间Hive表的迁移。

语法格式：

```
Export TABLE tablename TO 'filepath';
```

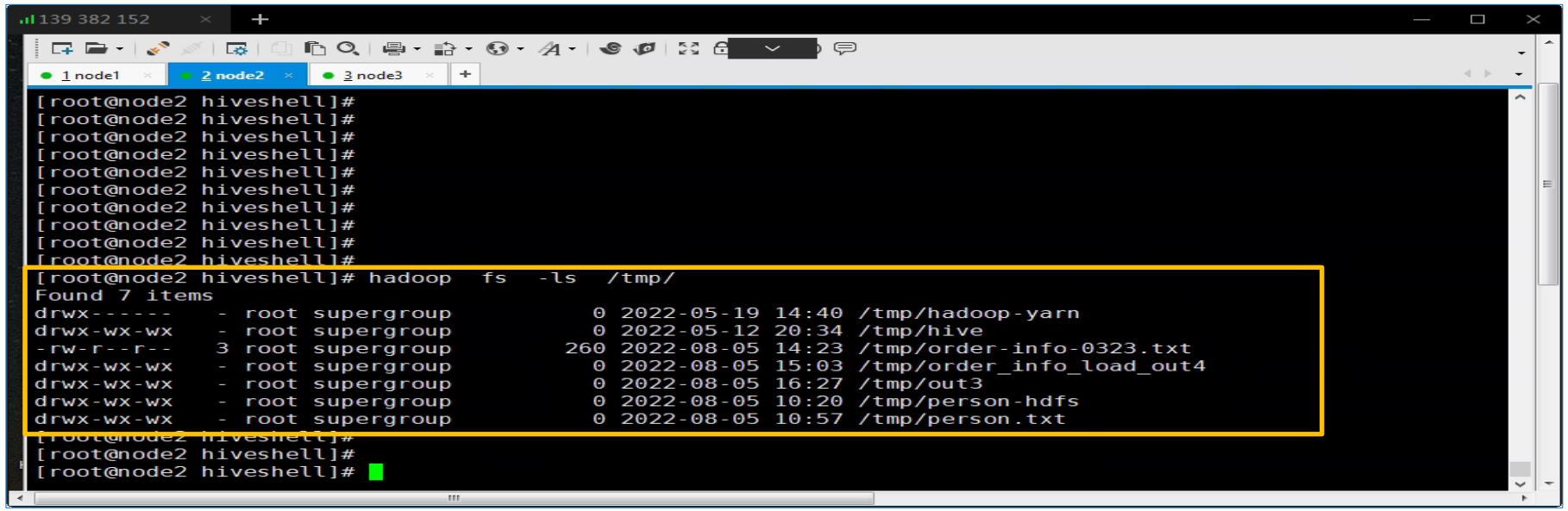
5.2.3 任务实施

例题 19：将表 order_info_load中的数据导出到 HDFS 文件系统中/tmp/order_info_load_out5 中。

操作过程：

step01：查看 HDFS 文件系统中/tmp 的目录结构查看 HDFS 文件系统中/tmp 的目录结构，如图 5-74 所示，其中没有/tmp/order_info_load_out5 目录。具体操作命令如下：

[root@node2 ~]# hdfs dfs -ls /tmp/



step02：　　将文件导出到 HDFS 相应目录中。

首先，将文件导出到相应的目录中，本题目中所使用目录为 HDFS 上的文件系统 /tmp/order_info_load_out5 目录。

具体操作命令如下：

```
Hive> Export table order_info_load to '/tmp/order_info_load_out5';
```

5.2.3 任务实施

step03：再次查看 HDFS 文件系统中/tmp 的目录结构。导出命名之后，再次查看 HDFS 文件系统中/tmp 的目录结构，如图 5-75 所示。具体操作命令如下：[root@node2 ~]# hadoopf s- ls /tmp/

由结果可知，其中/tmp/order_info_load_out5 已经存在。说明，该目录是在导出操作执行时自动生成的。

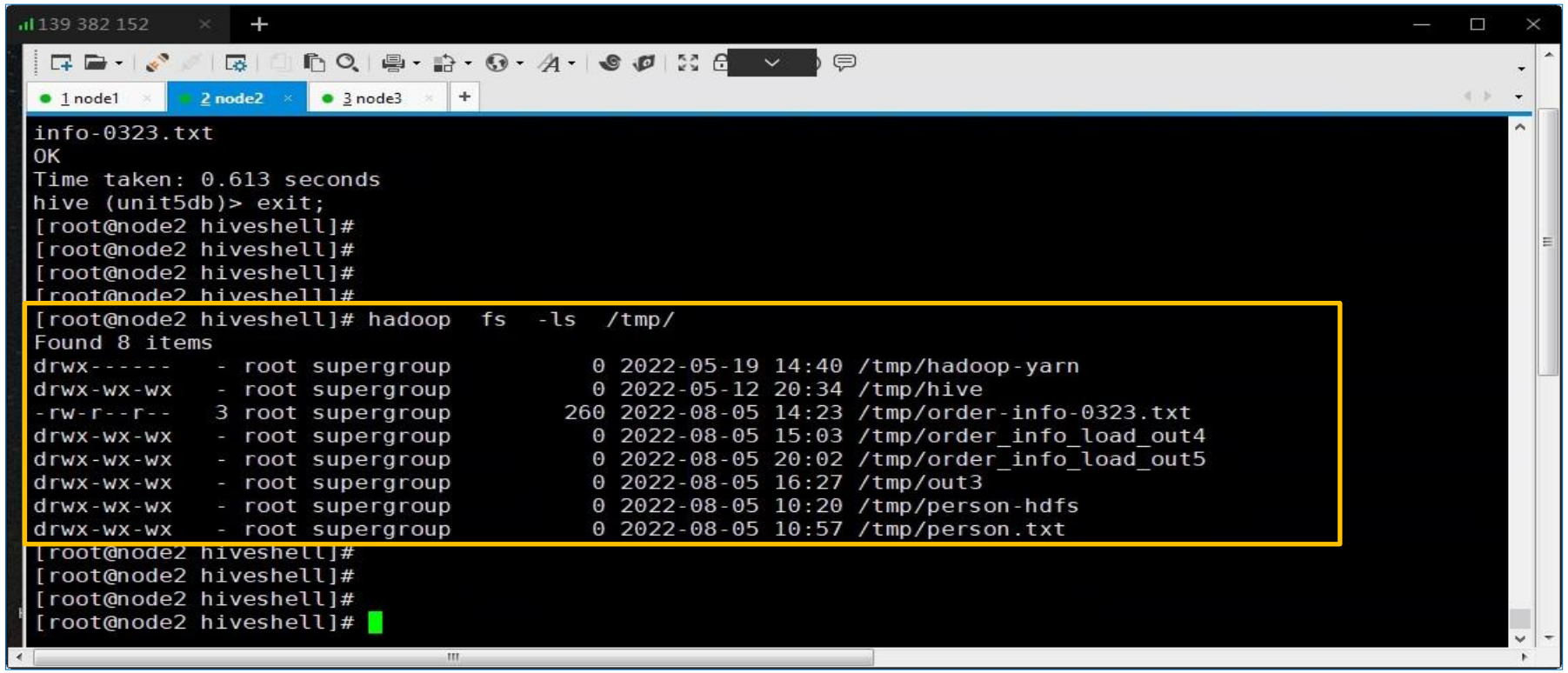


图 5-75 再次查看 HDFS 文件系统中/tmp 目录的结构

5.2.3 任务实施

step04：验证 HDFS 文件系统中/tmp/order_info_load_out5 的目录内容 验证 HDFS 文件系统中/tmp/order_info_load_out5 的目录内容，此处可以使用HDFS 命令进行验证，如图 5-76 所示，目录中存在内容。具体操作命令如下：

[root@node2 ~]# hadoop fs -ls /tmp/order_info_load_out5

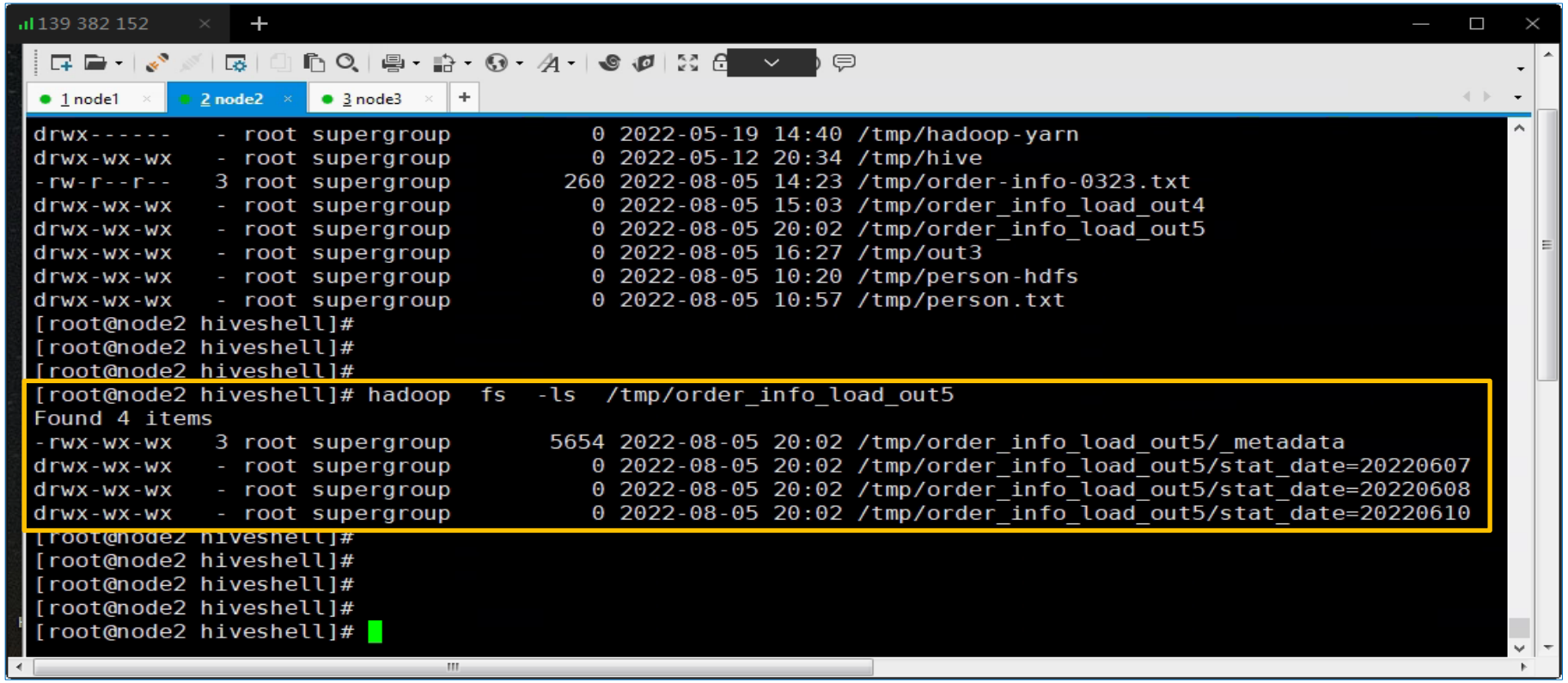
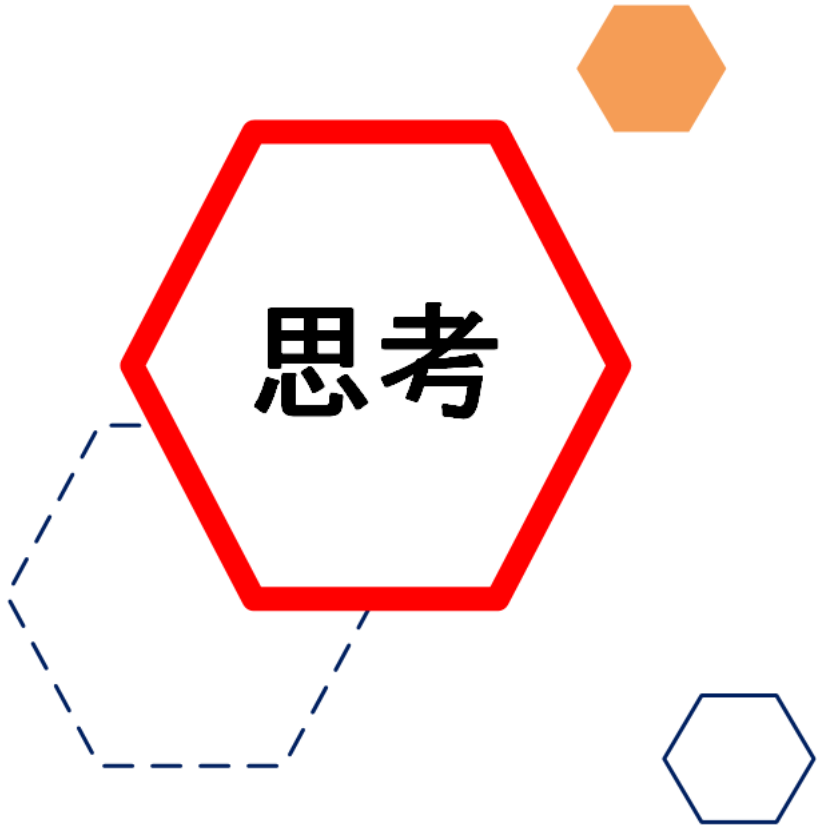


图 5-76 查看 HDFS 文件系统中/tmp/order_info_load_out5 目录的结构

5.2.3 任务实施

如果，该题目中读者需要验证/tmp/order_info_load_out5 中的具体数据，那么请参照例题 13 的方法，将 HDFS 文件系统中/tmp/order_info_load_out5 的数据下载到本地文件系统，然后利用 Linux 的相关命令打开文件，进行进一步验证即可。因为本题目旨在说明 Export 的导出功能，故而验证过程没有采用该方法，请感兴趣的读者，自行验证。

```
hadoop@hadoop01:/usr/datas/hiveshell$ hdfs dfs -cat /tmp/order_info_load_out5/stat_date=20260505/order-info-0208.txt
1,待支付,2026-02-08 12:20:11,2026-02-08 12:20:11
2,待支付,2026-02-08 14:20:11,2026 02-08 14:20:11
3,待支付,2026-02-08 16:20:11,2026-02-08 16:20:11
hadoop@hadoop01:/usr/datas/hiveshell$
```



适用场景：

1)这种方式只能导出数据到HDFS 上，但是导出的速度比较快，因此适用于 hive数据的批量迁移。

任务 5-3

本章小结

5.3.1 任务描述

我们在本章中首先学习了Hive生态系统中加载数据到表中的六种方式，

其中，重点介绍了通过Load加载数据、通过查询加载数据、查询语句中创建表并加载数据等三种方式；

然后，本章还详细介绍了Hive生态系统中导出数据的四种方式，

其中，Insert导出、Hadoop命令导出、Export命令导出等三种方式进行了重点说明。

在本章中，不仅理解和掌握Hive生态系统中加载数据到表中的几种方式，

同时理解和掌握Hive生态系统中导出数据的几种方式，

是日常生活中应用Hive解决实际问题的重要基础。

感谢您的观看

