Slide 70 — 任务5-2 数据导出（章节封面）

好的同学们，学完了数据导入，我们现在来学数据导出。

数据导出，英文叫 Data Export。

就是把 Hive 表里面的数据取出来，导出到本地文件系统或者 HDFS 上。

Slide 71 — 5.2.1 任务描述

在 Hive 的日常使用中，经常需要把表中的数据导出来。

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（xūqiú），

如果随意（suíyì）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（zīyuán）不必要的浪费（làngfèi）。

所以我们需要根据不同的场景选择最合适（héshì）的导出方式。

Slide 72 — 5.2.1 任务准备知识

从 Hive 中导出数据，实际上就是从 HDFS 中导出数据。

因为 Hive 的存储依赖（yīlài）于 HDFS 分布式文件系统。

语法中的关键字在前面已经讲过了，这里就不重复了。

Slide 73 — 5.2.3 数据导出方式总览

Hive 数据导出主要有四种方式：

第一种：Insert 导出 —— 通过查询语句导出

第二种：Hadoop 命令导出 —— 用 hdfs dfs -get 命令

第三种：Hive Shell 命令导出 —— 用 hive -e/-f 命令

第四种：Export 命令导出 —— 用于集群间迁移

下面我们一个一个来学习。

Slide 73 — 方式一：Insert 语句导出

第一种导出方式是用 INSERT 语句导出。

这种方式是把查询结果写入到文件系统中。

可以导出到本地文件系统，也可以导出到 HDFS。

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

Slide 74-78 — 例题14：导出到本地文件系统

例题14要求把 person 表中的数据导出到本地 /usr/datas/out1 目录中。

Step 01：查看本地目录

【实习命令 / 实践操作】查看本地目录结构

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

查看本地目录，确认 out1 目录还不存在。

Step 02：查看源表数据

【实习命令 / 实践操作】查看person表

```
hive> SELECT * FROM person;
```

person 表中共有80条记录。

Step 03：执行导出命令

【实习命令 / 实践操作】Insert导出到本地

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

注意关键字：

INSERT OVERWRITE —— 表示覆盖方式写入

LOCAL DIRECTORY —— 表示导出到本地目录（如果没有 LOCAL 就是导出到 HDFS）

Step 04：验证导出结果

【实习命令 / 实践操作】验证导出结果

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

【图片说明】图5-49至5-54展示了完整的导出过程。先查看了本地目录（out1不存在），然后执行了INSERT OVERWRITE LOCAL DIRECTORY命令，最后验证out1目录已自动生成，里面有80条导出的数据。

Slide 79-81 — 例题15：格式化导出到本地

例题15和例题14类似，但这次我们要按照格式化（géshìhuà）标准导出。

也就是说，我们要指定字段（zìduàn）之间的分隔符（fēngéfú）。

【实习命令 / 实践操作】格式化导出

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

这里多了 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' 这部分，

意思是用逗号（dòuhào）作为字段的分隔符。

这样导出的文件就是 CSV 格式，更容易阅读（yuèdú）和使用。

【实习命令 / 实践操作】验证格式化导出结果

```
[root@node2 datas]# cd /usr/datas/out2
[root@node2 out2]# ll
[root@node2 out2]# cat 000000_0
```

【图片说明】图5-55至5-58展示了格式化导出的过程和结果。导出后的数据以逗号分隔，整齐排列，共80条数据。

Slide 82-86 — 例题16：导出到 HDFS 文件系统

例题16要求把 person 表的数据导出到 HDFS 的 /tmp/out3 目录中。

和导出到本地的区别就是：去掉 LOCAL 关键字。

Step 01-02：查看 HDFS 目录和源表数据

【实习命令 / 实践操作】查看HDFS目录

```
[root@node2 ~]# hadoop fs -ls /tmp/
hive> SELECT * FROM person;
```

Step 03：导出到 HDFS

【实习命令 / 实践操作】Insert导出到HDFS

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

注意：这里没有 LOCAL，直接写 DIRECTORY，表示导出到 HDFS。

Step 04：验证导出结果

【实习命令 / 实践操作】验证HDFS导出结果

```
[root@node2 ~]# hadoop fs -ls /tmp/out3
[root@node2 ~]# cd /usr/datas/
[root@node2 datas]# hdfs dfs -get /tmp/out3/
[root@node2 datas]# cd out3
[root@node2 out3]# cat 000000_0
```

【图片说明】图5-59至5-63展示了导出到HDFS的完整过程，包括查看HDFS目录、执行导出命令、下载到本地验证。共导出80条数据。

Slide 87 — Insert 导出使用场景

Insert 导出方式有什么特点呢?

第一，这种方式会启动 MapReduce 任务来完成导出。
所以不适合导出很少的数据，因为即使数据量很小也会消耗（xiāohào）集群资源。

第二，适合对导出数据格式有要求的场景。
比如你可以指定数据文件的类型和字段分隔符。

第三，导出的路径只能指定到文件夹，不能指定到某个具体的文件。
而且文件夹下面可能会产生多个文件，不一定只有一个。

Slide 88 — 方式二(방법2)：Hadoop 命令导出

第二种导出方式是用 Hadoop 命令来导出。
这种方式很直接（zhíjiē），就是用 hdfs dfs -get 命令把 HDFS 上的文件下载到本地。

语法格式非常简单：

【实习命令 / 实践操作】Hadoop命令导出语法

```
hdfs dfs -get [HDFS文件目录] [本地目录]
```

就像从网上下载（xiàzài）文件一样，-get 就是“获取”的意思。

Slide 89-95 — 例题17：用 Hadoop 命令导出数据

例题17要求把 HDFS 上 /tmp/out3 中的数据导出到本地 /usr/datas/hadoop 目录。

Step 01：确认 HDFS 数据

【实习命令 / 实践操作】查看HDFS目录

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

Step 02：创建本地目录并下载

【实习命令 / 实践操作】创建本地目录

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

【实习命令 / 实践操作】下载HDFS数据到本地

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

Step 03：验证数据

【实习命令 / 实践操作】验证下载的数据

```
[root@node2 datas]# cd /usr/datas/hadoop
[root@node2 hadoop]# ll
[root@node2 hadoop]# cd out3
[root@node2 out3]# cat 000000_0
```

【图片说明】图5-64至5-70展示了完整的操作过程。先确认HDFS上 /tmp/out3 数据存在，然后创建了本地目录 /usr/datas/hadoop，用 hdfs dfs -get 命令下载数据，最后验证数据共80条，下载成功。

Slide 96 — 方式三：Hive Shell 命令导出

第三种方式是用 Hive Shell 命令导出。

也就是直接在 Linux 命令行中使用 hive 命令，加上参数（cānshù） -e 或者 -f 来导出数据。

语法格式：

【实习命令 / 实践操作】Hive Shell导出语法

```
hive -e "SQL语句" > filepath
hive -f SQL文件 > filepath
```

-e 后面直接跟 SQL 语句。

-f 后面跟一个包含 SQL 语句的文件。

> 是 Linux 的重定向（chóngdìngxiàng）符号，把查询结果输出到文件中。

Slide 97-101 — 例题18：Hive Shell 导出数据

例题18要求把 order_info_load 表的数据导出到 /usr/datas/hiveshell 目录。

Step 01：查看源表数据

【实习命令 / 实践操作】查看源表数据

```
hive> SELECT * FROM order_info_load;
```

表中共有13条数据。

Step 02：创建本地目录

【实习命令 / 实践操作】创建导出目录

```
[root@node2 datas]# cd /usr/datas
[root@node2 datas]# mkdir hiveshell
```

Step 03：执行 Hive Shell 导出

【实习命令 / 实践操作】Hive Shell -e 导出

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

注意：这条命令是在 Linux 命令行执行的，不是在 Hive 里面！

因为这里用的是数据库名.表名的格式（unit5db.order_info_load），

所以不需要先 USE unit5db。

【实习命令 / 实践操作】验证导出结果

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

【图片说明】图5-71至5-73展示了完整的操作过程，包括查看源表数据、创建目录、执行导出命令和验证结果。导出文件order_info_load_out.txt中包含了13条完整数据。

Slide 102 — Hive Shell 导出使用场景

Hive Shell 导出方式的特点：

第一，不会启动 MapReduce 任务，所以速度快，适合数据量较少的情况。

第二，支持指定导出的文件名，而且导出结果只有一个文件。

第三，只能导出到本地文件系统，不能导出到 HDFS。

第四，不支持指定导出文件的分隔符。

Slide 103 — 方式四(방식 4)：Export 语句导出

第四种方式是用 Export 语句导出。

前面我们在 Import 导入的时候提到过，Export 和 Import 是配对使用的。

主要用于两个 Hadoop 集群之间的 Hive 表迁移（qiānyí）。

语法格式：

【实习命令 / 实践操作】Export导出语法

```
hive> EXPORT TABLE tablename TO 'filepath';
```

Slide 104-108 — 例题19：Export 导出数据

例题19要求把 order_info_load 表的数据导出到 HDFS 的 /tmp/order_info_load_out5 目录。

Step 01：查看 HDFS 目录

【实习命令 / 实践操作】查看HDFS目录

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

确认 /tmp/order_info_load_out5 目录还不存在。

Step 02：执行 Export 导出

【实习命令 / 实践操作】Export导出

```
hive> EXPORT TABLE order_info_load TO '/tmp/order_info_load_out5';
```

Step 03-04：验证导出结果

【实习命令 / 实践操作】验证导出结果

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

【图片说明】图5-74至5-76展示了Export导出的完整过程。导出前 /tmp 目录下没有 order_info_load_out5，执行Export后该目录自动生成，里面包含了导出的数据文件。

Slide 109 — Export 导出使用场景

Export 导出方式的特点：

第一，只能导出数据到 HDFS 上，不能导出到本地。

第二，导出速度比较快。

第三，主要适用于 Hive 数据的批量迁移（pīliàng qiānyí）。

简单记忆：Export 是打包搬走，适合整体迁移；Insert 导出是选择性提取，更灵活。

Slide 110 — 任务5-3 本章小结（章节封面）

好了同学们，到这里我们今天的主要内容就全部讲完了。

现在让我们来做一个总结。

Slide 111 — 5.3.1 本章小结内容

在这一章中，我们学了两大部分内容：

第一部分：数据导入（五种方式）

① 直接向表中加载数据 —— INSERT INTO ... VALUES

② 通过 Load 加载数据 —— LOAD DATA [LOCAL] INPATH ... INTO TABLE ...（重点！）

③ 通过查询加载数据 —— INSERT [OVERWRITE/INTO] TABLE ... SELECT ...（重点！）

④ 查询语句中创建表并加载数据 —— CREATE TABLE ... AS SELECT ...（重点！）

⑤ Import 导入 Hive 表数据 —— IMPORT TABLE ... FROM ...

第二部分：数据导出（四种方式）

① Insert 导出 —— INSERT OVERWRITE [LOCAL] DIRECTORY ...（重点！）

② Hadoop 命令导出 —— hdfs dfs -get ...（重点！）

③ Hive Shell 命令导出 —— hive -e/-f ... > filepath

④ Export 命令导出 —— EXPORT TABLE ... TO ...（重点！）

掌握（zhǎngwò）这些数据导入和导出的方法，

是我们日常使用 Hive 解决实际问题的重要基础（jīchǔ）。

💡 建议（jiànyì）同学们课后多练习，特别是 LOAD DATA 和 INSERT 相关的操作。

Slide 112 — 感谢您的观看

好的，同学们，今天的课就到这里了。

今天我们学了 Hive 数据操作中最核心的两个内容：数据导入和数据导出。

一共学了九种方法（五种导入 + 四种导出）。

课后请大家：

第一，回顾今天的笔记和讲义，把每种方式的语法格式记住。

第二，自己动手做一遍今天课上的例题。

第三，完成任务5-4的本章习题。

如果有任何问题，欢迎随时来问我。

谢谢大家！下课！