任务 3-4 存储数据

【幻灯片 1-3】 任务3-4 存储数据 - 任务描述

最后一个任务——存储数据。

다음 슬라이드 넘기면서]

为什么要了解数据存储?

因为选对存储格式（gé shì），即使不改SQL，性能也能有数量级（shù liàng jí）的提升!

Hive建表后，有两个重要的存储位置（wèi zhì）：

1. 表的元数据（yuán shù jù）—— 存在关系型数据库中（比如MySQL），记录表的结构信息。

2. 表的数据内容 —— 以文件形式存在HDFS中。

[다음 슬라이드 넘기면서]

Hive的表结构包括：普通表（Table）、外部表（External）、分区表（Partition）、桶表（Bucket）。

这些我们后面的课程会详细讲解（xiáng xì jiǎng jiě）。

【幻灯片 4】 3.4.2 Hive表存储格式

Hive支持多种存储格式：

1. TextFile（文本格式）—— 最简单，就是普通的文本文件。默认格式。

2. SequenceFile（序列文件）—— Hadoop的二进制格式。

3. RCFile（行列式文件）—— 结合了行存储和列存储的优点。

4. ORC（优化的行列式文件）—— 目前最推荐的格式，压缩（yā suō）效果好，查询速度快。

【幻灯片 5】 3.4.3 行式存储和列式存储

数据存储有两种基本方式：行式存储和列式存储。

行式存储（háng shì cún chǔ）：

把一整行数据存在一起。就像Excel里一行一行地存。

优点：查询一整行数据很快。

缺点：如果只需要某一列的数据，也要读取整行。

列式存储（liè shì cún chǔ）：

把同一列的数据存在一起。

优点：只查询某几列时非常快，而且同一列的数据类型相同，压缩效果更好。

缺点：查询一整行数据时需要拼接（pīn jiē）多个列。

> 💡 *数据仓库通常用列式存储（如ORC格式），因为分析查询往往只涉及几个字段。*

【幻灯片 6】 3.4.3 文本文件格式 TextFile

TextFile是Hive默认（mò rèn）的存储格式。

特点：每一行是一条记录，每行以换行符结尾。

优点：简单直观，可以直接用cat命令查看。

缺点：不做压缩，占用空间大，解析速度慢。

虽然可以配合Gzip、Bzip2压缩使用，但压缩后Hive不能对数据进行切分（qiē fēn），也就无法并行处理。

TextFile主要用在数据源层（ODS层），也就是最原始（yuán shǐ）数据的存储。

【幻灯片 7-8】 3.4.3 顺序文件格式 SequenceFile

SequenceFile是Hadoop API提供的一种二进制文件格式。

它用<key, value>键值对的形式存储数据。

有三种压缩（yā suō）类型：

1. 无压缩（None）—— 不压缩，速度快但占空间。

2. 记录压缩（Record Compression）—— 压缩每一条记录。

3. 块压缩（Block Compression）—— 一次性压缩多条记录，压缩效果最好。

课程总结（kè chéng zǒng jié）

好，今天的课程到这里就结束了。我们来回顾（huí gù）一下今天学了什么：

任务3-1 认识数据类型：

- 原生类型：数值（INT/DOUBLE）、字符串（STRING）、日期（DATE）、布尔（BOOLEAN）
- 复杂类型：ARRAY数组、MAP映射、STRUCT结构体
- 类型转换：隐式转换（小→大自动转）、显式转换（CAST函数）

任务3-2 认识运算符：

- 关系运算符：=, <>, >, <, >=, <=, IS NULL, LIKE, RLIKE
- 算术运算符：+, -, *, /, %
- 逻辑运算符：AND, OR, NOT
- 复杂运算符：A[n], M['key'], S.x

任务3-3 认识Hive命令：

- 两代客户端：Hive CLI 和 Beeline CLI
- HDFS命令、Linux命令、Tab补全、历史命令

任务3-4 存储数据：

- 行式存储 vs 列式存储
- TextFile（默认）、SequenceFile、RCFile、ORC

大家课后要多实践（shí jiàn），把今天学的命令都敲（qiāo）一遍。有问题随时问！

谢谢大家！下课！