任务 3-4 存储数据

3.4.1 任务描述

存储Hive数据，是Hive操作数据的基础。

选择一个合适的底层数据存储文件格式，

即使在不改变当前Hive SQL的情况下，性能也能得到数量级的提升。

Hive建表后，表的元数据存储在关系型数据库中，

表的数据内容以文件形式存储在Hadoop集群的HDFS中。







1．Hive数据存储

Hive的数据存储其实是在HDFS文件系统中的，在Hive数据仓库中，表中的数据没有专门的数据存储格式，也没有为数据建立索引，用户可以自由地组织Hive表中的数据，可以使用文本文件存储表中的数据，也可使用CSV文件存储表中的数据。

Hive的数据都存储在HDFS，Hive的结构包括表（Table）、外部表（External）、分区表（Partition）、桶表（Bucket）。

2．Hive表存储格式

Hive支持表的存储格式主要有

TextFile（文本格式）、

RCFile（行列式文件）、

SequenceFile（二进制序列化文件）、

ORC（有好的行列式文件）

3.4.3 任务实施

1．行式存储和列式存储

行存储模式就是把一整行存在一起，包含所有的列，这是最常见的模式。查询满足条件的一整行数据时，行式存储只需要找到其中一个值，其余的值都在相邻地方。

列式存储结构，相对于行式存储来说，新兴的Hbase、HP Vertica、EMC Greenplum等分布式数据库均采用列式存储。

在列式存储中，数据是按照列为基础逻辑单元进行存储的，一列中的数据在存储介质中以连续形式存在。

2．文本文件格式

文本文件（TextFile）格式是Hive在创建表时默认的存储格式，每一行都是一条记录，每行都以换行符（\n）结尾。

数据不做压缩，磁盘开销大，数据解析开销大。可结合Gzip、Bzip2使用(系统自动检查，执行查询时自动解压)，但使用这种方式，hive不会对数据进行切分，从而无法对数据进行并行操作。

文本文件主要使用场景在数据源层ODS层，针对需要使用脚本load加载数据到Hive数据仓库表中情况。

3．顺序文件格式

顺序文件（SequenceFile）格式是Hadoop API提供的一种二进制文件支持，是Hadoop用来存储二进制形式的<key,value>键值对二设计的一种文件存储结构。

（1）无压缩类型（None）

（2）记录压缩类型 (Record Compression)

（3）块压缩类型 (Block Compression)

3．顺序文件格式

顺序文件（SequenceFile）格式是Hadoop API提供的一种二进制文件支持，是Hadoop用来存储二进制形式的<key,value>键值对二设计的一种文件存储结构。

（1）无压缩类型（None）

（2）记录压缩类型 (Record Compression)

（3）块压缩类型 (Block Compression)

感谢您的观看



《Hive数据仓库应用教程》