项目四 Hive数据定义



《Hive数据仓库应用教程》

目 录 / CONTENTS

1 任务4-1 HiveQL操作数据库

2 任务4-2 HiveQL操作数据表

3 任务4-3 HiveQL操作分区表

4 任务4-4 HiveQL操作分桶表

5 任务4-5 HiveQL操作视图

6 任务4-6 HiveQL相关操作

任务 4-4 HiveQL操作分桶表

4.4.1 任务描述

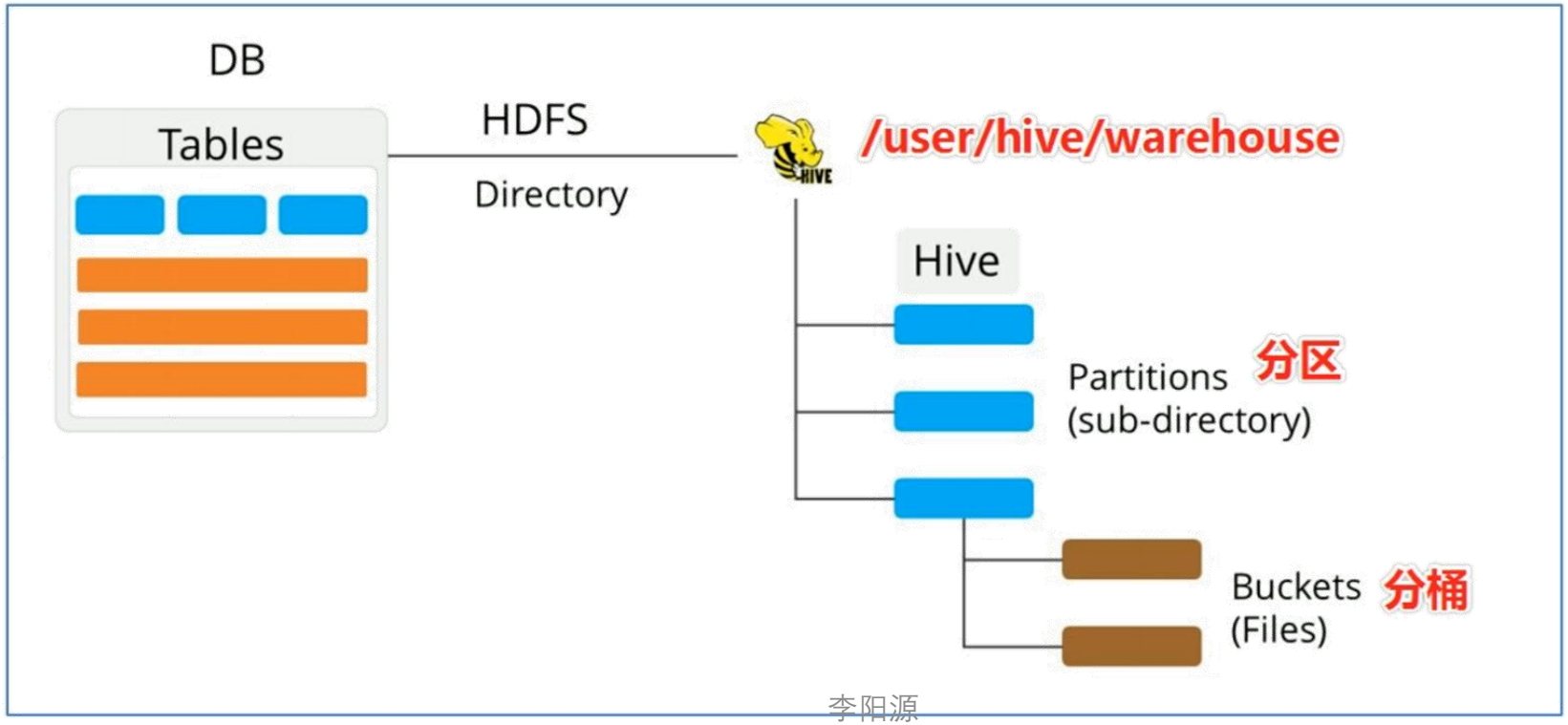
Hive的分区可以将整体数据划分成多个分区，从而优化查询，

但是并非所有的数据都可以被合理的分区，会出现每个分区数据大小不一致的问题，有的分区数据量很大，有的分区数据量确很小，这就是我们常说的数据倾斜。

为了解决分区可能带来的数据倾斜问题，Hive提供了分桶技术，Hive中的分桶是指定分桶表的某一列，让该列数据按照哈希取模的方式随机、均匀的分发到各个桶文件中。

概念

- 分桶表也称作桶表，名称来源于建表语法中的 bucket 单词，是一种用于优化查询而设计的表类型。
- 分桶表对应的数据文件在底层会被分解为多个部分，通俗来说就是被拆分成多个独立的小文件。
- 在分桶时，需要指定根据哪个字段将数据分为几桶（即几个部分）。



4.4.2 任务准备知识

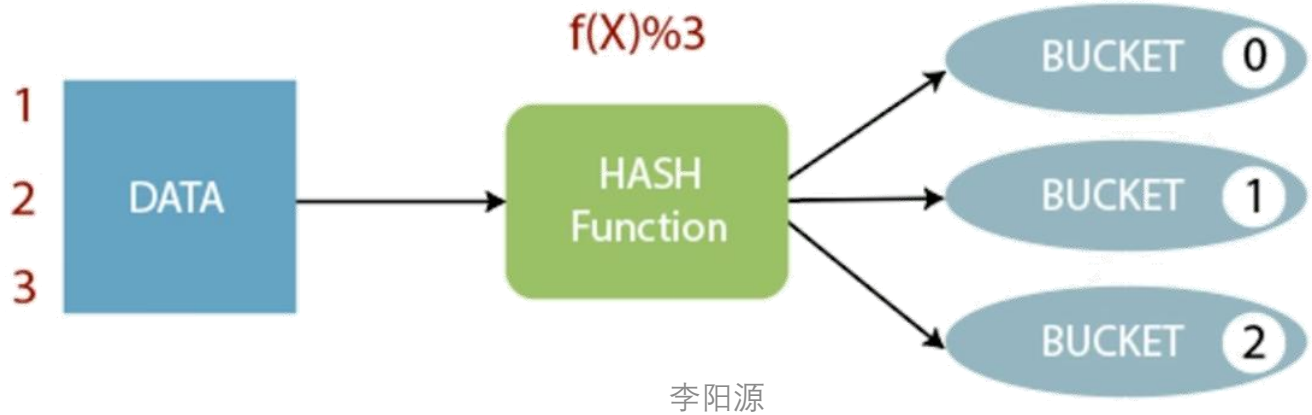
1．规则

- 分桶规则如下：桶编号相同的数据会被分到同一个桶当中。

```
Bucket number = hash_function(bucketing_column) mod num_buckets
分桶编号      = 哈希方法（分桶字段）              取模  分桶个数
```

- hash_function 取决于分桶字段 bucketing_column 的类型：

1. 如果是 int 类型，hash_function(int) == int；
2. 如果是其他类型，如 bigint、string 或复杂数据类型，hash_function **比较棘手**，将是从该类型派生的某个数字，比如 hashcode 值。



4.4.2 任务准备知识

1．规则

- 分桶规则如下：桶编号相同的数据会被分到同一个桶当中。

```
Bucket number = hash_function(bucketing_column) mod num_buckets
分桶编号       = 哈希方法（分桶字段）              取模  分桶个数
```

- hash_function 取决于分桶字段 bucketing_column 的类型：

1. 如果是 int 类型，hash_function(int) == int；
2. 如果是其他类型，如 bigint、string 或复杂数据类型，hash_function **比较棘手**，将是从该类型派生的某个数字，比如 hashcode 值。

2．完整语法树

- **CLUSTERED BY (col_name)** 表示根据哪个字段进行分桶；
- **INTO N BUCKETS** 表示分为几桶（即分成几个部分）。
- 需要注意的是，**分桶的字段必须是表中已经存在的字段**。

```
--分桶表建表语句
CREATE [EXTERNAL] TABLE [db_name.]table_name
[(col_name data_type, ...)]
CLUSTERED BY (col_name)
INTO N BUCKETS;
```

4.4.3 任务实施

分桶表的创建

- 现有美国2021-1-28号，各个县county的新冠疫情累计案例信息，包括确诊病例和死亡病例，数据格式如下所示；
- 字段含义：count_date（统计日期），county（县），state（州），fips（县编码code），cases（累计确诊病例），deaths（累计死亡病例）。

us-covid19-counties.dat

```
2021-01-28,Jefferson,Alabama,01073,65992,1101
2021-01-28,Lamar,Alabama,01075,1230,24
2021-01-28,Lauderdale,Alabama,01077,8088,127
2021-01-28,Lawrence,Alabama,01079,2616,61
2021-01-28,Lee,Alabama,01081,13602,113
2021-01-28,Limestone,Alabama,01083,8564,89
2021-01-28,Lowndes,Alabama,01085,1195,38
2021-01-28,Macon,Alabama,01087,1260,38
2021-01-28,Madison,Alabama,01089,29098,248
2021-01-28,Marengo,Alabama,01091,2152,34
```

4.4.3 任务实施

分桶表的创建

- 根据state州把数据分为5桶，建表语句如下：

```
CREATE TABLE itcast.t_usa_covid19_bucket(
    count_date string,
    county string,
    state string,
    fips int,
    cases int,
    deaths int)
CLUSTERED BY(state) INTO 5 BUCKETS;--分桶的字段一定要是表中已经存在的字段
```

- 在创建分桶表时，还可以指定分桶内的数据排序规则：

```
CREATE TABLE itcast.t_usa_covid19_bucket_sort(
    count_date string,
    county string,
    state string,
    fips int,
    cases int,
    deaths int)
CLUSTERED BY(state)
sorted by (cases desc) INTO 5 BUCKETS;--指定每个分桶内部根据cases倒序排序
```

4.4.3 任务实施

分桶表的数据加载

```
-- 分桶表的数据加载
-- step1: 开启分桶功能（从 Hive 2.0 开始不再需要显式设置）
-- set hive.enforce.bucketing=true;

-- step2: 把源数据加载到普通 Hive 表中
drop table if exists itcast.t_usa_covid19;
hdfs dfs -rm -r /user/hive/warehouse/itcast.db/t_usa_covid19 --위 명령으로 삭제 안되는 경우

CREATE TABLE itcast.t_usa_covid19(
    count_date string,
    county string,
    state string,
    fips int,
    cases int,
    deaths int)
row format delimited fields terminated by ',';

-- 将源数据上传到 HDFS t_usa_covid19 表对应的路径下
hdfs dfs -put us-covid19-counties.dat /user/hive/warehouse/itcast.db/t_usa_covid19

-- step3: 使用 insert+select 语法将数据加载到分桶表中
insert into t_usa_covid19_bucket select * from t_usa_covid19; --시간 매우 많이 걸림. 기다릴것
```

4.4.3 任务实施

分桶表的数据加载

Browse Directory

/user/hive/warehouse/itcast.db/t_usa_covid19_bucket Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	19.54 KB	Feb 06 16:51	1	128 MB	000000_0
-rw-r--r--	hadoop	supergroup	32.92 KB	Feb 06 16:51	1	128 MB	000001_0
-rw-r--r--	hadoop	supergroup	45.48 KB	Feb 06 16:51	1	128 MB	000002_0
-rw-r--r--	hadoop	supergroup	11.36 KB	Feb 06 16:51	1	128 MB	000003_0
-rw-r--r--	hadoop	supergroup	24.2 KB	Feb 06 16:51	1	128 MB	000004_0

Showing 1 to 5 of 5 entries Previous 1 Next

4.4.4 Hive Bucketed Tables 分桶表使用好处

- （1/3）基于分桶字段查询时，减少全表扫描

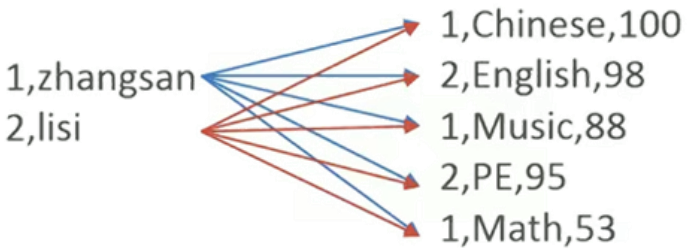
```
--基于分桶字段state查询来自于New York州的数据
--不再需要进行全表扫描过滤
--根据分桶的规则hash_function(New York) mod 5计算出分桶编号
--查询指定分桶里面的数据 就可以找出结果 此时是分桶扫描而不是全表扫描
select *
from t_usa_covid19_bucket where state="New York";
```

4.4.4 Hive Bucketed Tables 分桶表使用好处

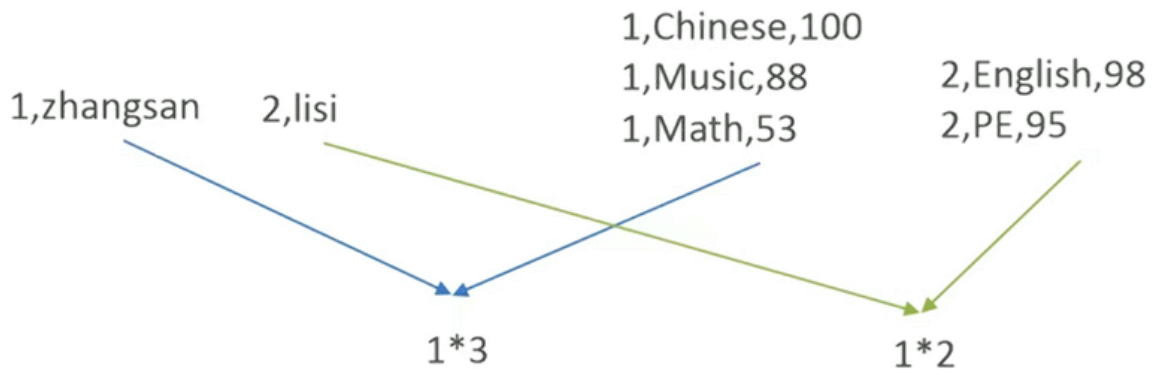
- （2/3）JOIN时可以提高MR程序效率，减少笛卡尔积数量

根据join的字段对表进行分桶操作（比如下图中id是join的字段）

```
a  join  b on a.id =b.id
```



需要关联2*5=10次，符合结果的是5个 效率50%



需要关联3+2=5次，符合结果的是5个 效率100%

4.4.4 Hive Bucketed Tables 分桶表使用好处

- （3/3）分桶表数据进行**高效抽样**

当数据量特别大时，对全体数据进行处理存在困难时，抽样就显得尤其重要了。抽样可以从被抽取的数据中估计和推断出整体的特性，是科学实验、质量检验、社会调查普遍采用的一种经济有效的工作和研究方法。

感谢您的观看



《Hive数据仓库应用教程》