2019

企业级数据仓库实战

Hive DML操作介绍

LOAD DATA [LOCAL] INPATH 'filepath' [OVERWRITE] INTO TABLE tablename [PARTITION (partcol1=val1, partcol2=val2 …)]
LOAD DATA [LOCAL] INPATH 'filepath' [OVERWRITE] INTO TABLE tablename [PARTITION (partcol1=val1, partcol2=val2 …)]
[INPUTFORMAT 'inputformat' SERDE 'serde'] (3.0 or later)

例子：
create table score2
         (
                    name string
    , score map<string,string>
)
PARTITIONED BY (pt STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
COLLECTION ITEMS TERMINATED BY ','
MAP KEYS TERMINATED BY ':';

LOAD DATA LOCAL INPATH '/Users/PycharmProjects/DW/File/score' OVERWRITE INTO TABLE score2 PARTITION (pt='20191111');

INSERT OVERWRITE TABLE tablename1 [PARTITION (partcol1=val1, partcol2=val2 ...) [IF NOT EXISTS]] select_statement1 FROM from_statement;
INSERT INTO TABLE tablename1 [PARTITION (partcol1=val1, partcol2=val2 ...)] select_statement1 FROM from_statement;

例子：
INSERT OVERWRITE TABLE score2 PARTITION(pt='20191112')
SELECT
    22 AS ID
    ,str_to_map('course:Math|score:100','\\|',':') AS course;

FROM from_statement
INSERT OVERWRITE TABLE tablename1 [PARTITION (partcol1=val1, partcol2=val2 ...) [IF NOT EXISTS]] select_statement1
[INSERT OVERWRITE TABLE tablename2 [PARTITION ... [IF NOT EXISTS]] select_statement2]
[INSERT INTO TABLE tablename2 [PARTITION ...] select_statement2] ...;
FROM from_statement
INSERT INTO TABLE tablename1 [PARTITION (partcol1=val1, partcol2=val2 ...)] select_statement1
[INSERT INTO TABLE tablename2 [PARTITION ...] select_statement2]
[INSERT OVERWRITE TABLE tablename2 [PARTITION ... [IF NOT EXISTS]] select_statement2] ...;
举例：
FROM score2
INSERT OVERWRITE TABLE score2 PARTITION(pt='20191113') SELECT name,score
INSERT OVERWRITE TABLE score3 PARTITION(pt='20191113') SELECT name,score;

数据动态分区写入

INSERT OVERWRITE TABLE tablename PARTITION (partcol1[=val1], partcol2[=val2] ...) select_statement FROM from_statement;
INSERT INTO TABLE tablename PARTITION (partcol1[=val1], partcol2[=val2] ...) select_statement FROM from_statement;

配置项	Default	Note
hive.error.on.empty.partition	FALSE	Whether to throw an exception if dynamic partition insert generates empty results
hive.exec.dynamic.partition	TRUE	Needs to be set to `true` to enable dynamic partition inserts
hive.exec.dynamic.partition.mode	strict	In `strict` mode, the user must specify at least one static partition in case the user accidentally overwrites all partitions, in `nonstrict` mode all partitions are allowed to be dynamic
hive.exec.max.created.files	100000	Maximum number of HDFS files created by all mappers/reducers in a MapReduce job
hive.exec.max.dynamic.partitions	1000	Maximum number of dynamic partitions allowed to be created in total
hive.exec.max.dynamic.partitions.pernode	100	Maximum number of dynamic partitions allowed to be created in each mapper/reducer node

举例：

```
set hive.exec.dynamic.partition = TRUE;
set hive.exec.dynamic.partition.mode= nonstrict;
EXPLAIN
INSERT OVERWRITE TABLE score2 PARTITION(pt)
SELECT
    name
    ,score
    ,pt
FROM score3;
;
```

脚本开发

表创建
外部表、ORC
\t分割、Location指定

UDF/临时表创建
删除UDF/临时表
创建UDF/临时表

临时表删除
清理临时表

01 **02** **03** **04** **05** **06**

指定库和队列
声明使用的库
指定本账号所属的队列

分区加载
删除分区
加载分区
指定分区Location

数据写入
通过覆盖的方式写入

```
—— 设置队列和参数
SET mapreduce.job.queuename=root.default;
SET mapreduce.map.memory.mb=8191;
—— 切换数据库
USE dw;
—— 建表
CREATE TABLE IF NOT EXISTS score2(name string, score map<string,string>)PARTITIONED BY (pt STRING);
—— 挂在分区
ALTER TABLE score2 DROP IF EXISTS PARTITION(pt='20191111');
ALTER TABLE score2 ADD IF NOT EXISTS PARTITION(pt='20191111');
—— 申明函数
CREATE FUNCTION default.function_name AS class_name USING JAR 'file_name';
—— 数据写入
INSERT OVERWRITE TABLE score2 PARTITION(pt='20191112')
SELECT
    22 AS ID
    ,str_to_map('course:Math|score:100','\\|',':') AS course;
—— 删除临时表(如果有)
DROP TABLE IF EXISTS tmp_score2_xxx;
```