2019

企业级数据仓库实战

Hive 查询语句

01

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
    FROM table_reference
    [WHERE where_condition]
    [GROUP BY col_list]
    [ORDER BY col_list]
    [CLUSTER BY col_list
        | [DISTRIBUTE BY col_list] [SORT BY col_list]
    ]
   [LIMIT [offset,] rows]
```

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：默认返回值即为ALL
　　　DISTINCT 表示对整行数据进行去重

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：FROM 后面跟数据库名.表名

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：WHERE 后面跟具体的条件

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
    FROM table_reference
    [WHERE where_condition]
    [GROUP BY col_list]
    [ORDER BY col_list]
    [CLUSTER BY col_list
        | [DISTRIBUTE BY col_list] [SORT BY col_list]
    ]
   [LIMIT [offset,] rows]
```

解释：GROUP BY 后面跟分组汇总的字段

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：ORDER BY 后面跟排序的字段
可以多个字段排序并制定排序规则 ASC代表升序
DESC 代表降序

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
   FROM table_reference
   [WHERE where_condition]
   [GROUP BY col_list]
   [ORDER BY col_list]
   [CLUSTER BY col_list
       | [DISTRIBUTE BY col_list] [SORT BY col_list]
   ]
  [LIMIT [offset,] rows]
```

解释：CLUSTER BY 控制Map端到Reduce如何划分

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：DISTRIBUTE BY 与CLUSTER BY 类似

ORDER BY 与 SORT BY 区别：
ORDER BY 为全局排序，最终会集中到一个Reduce，效率低
SORT BY 为部分排序，其只保证单个Reduce内是有序的
一般需要和DISTRIBUTE BY 与CLUSTER BY配合使用

Hive查询语句

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

解释：LIMIT 限制最终查询的数量

Hive 子查询语句

```
SELECT col
FROM (
  SELECT a+b AS col
  FROM t1
) t2
```

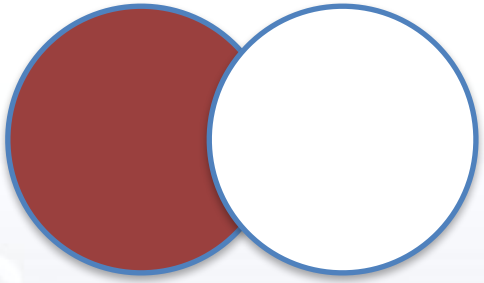
解释：即FROM后面跟的是另外一个SELECT语句

Hive join

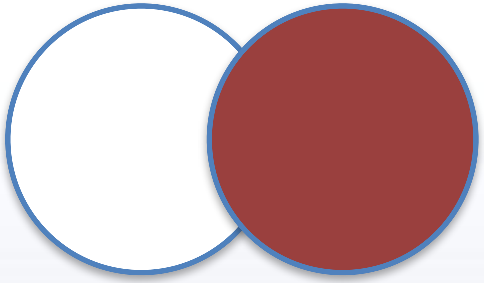
```
join_table:
    table_reference [INNER] JOIN table_factor [join_condition]
  | table_reference {LEFT|RIGHT|FULL} [OUTER] JOIN table_reference join_condition
  | table_reference LEFT SEMI JOIN table_reference join_condition
  | table_reference CROSS JOIN table_reference [join_condition] (as of Hive 0.10)

table_reference:
    table_factor
  | join_table

table_factor:
    tbl_name [alias]
  | table_subquery alias
  | ( table_references )

join_condition:
    ON expression
```

解释：即FROM后面跟的是另外一个SELECT语句

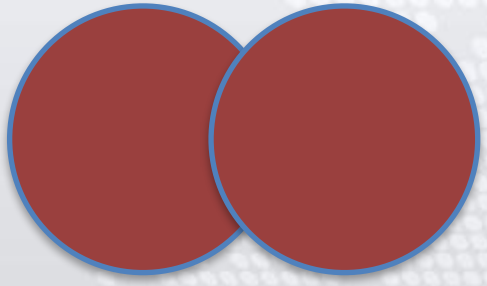
Hive join



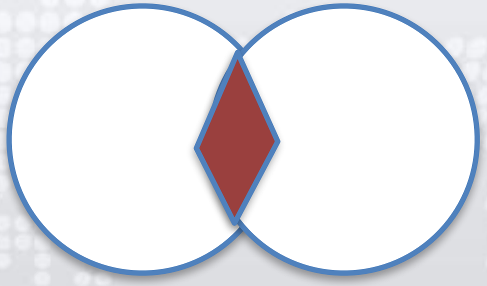
LEFT [OUTER] JOIN



RIGHT [OUTER] JOIN



FULL (OUTER) JOIN



INNER JOIN

Hive join 两种写法

写法一：

```
SELECT *
FROM table1 t1, table2 t2, table3 t3
WHERE t1.id = t2.id AND t2.id = t3.id AND t1.zipcode = ‘02535';
```

写法二：

```
SELECT k1, v1, k2, v2
FROM a JOIN b ON k1 = k2;
```

思考题

1、order by 与 sort by 有何区别?

2、distinct去重可以使用那种方式进行改写?

3、常见的join有哪几种，请分别解释区别

4、hive中排序为何会比较耗费时间

5、hive 如何实现not join

THANK YOU FOR YOUR GUIDANCE.

谢谢