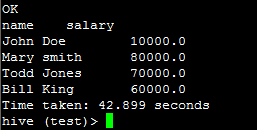
Hadoop学习之Hive编程指南学习(五)

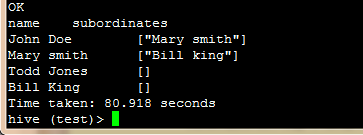
1. HiveQL的查询
   1. Select … FROM 语句
      1. Hive> select name,salary from emp;

select e.name,e.salary from emp e;两条语句等价，但在设计不同表时很有用。



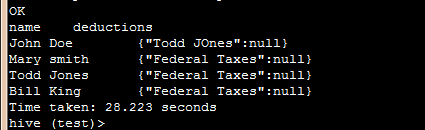
如果选择的列是个集合，hvie会使用json语法输出，是数组的话值会被括在[…]内，逗号分隔的列表中，集合的字符串元素加引号，而string类型的劣质不加引号

* + 1. hive (test)> select name,subordinates from emp;



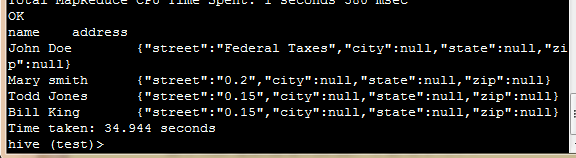
Deductions列是个map使用json表达{…}内以都好分隔的 “键：值”列进行表示。

* + 1. hive (test)> select name,deductions from emp;



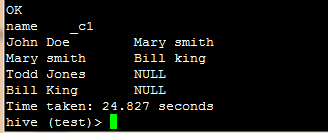
显示null是数据格式原因，不影响学习。

Address列是struct，使用json map格式表示



引用集合数据类型中的元素

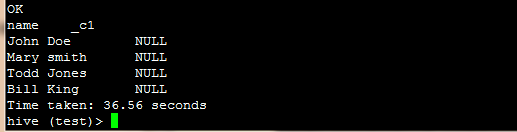
* + 1. hive (test)> select name,subordinates[0] from emp;



不存在该元素就返回null，提取出的string将不加引号。

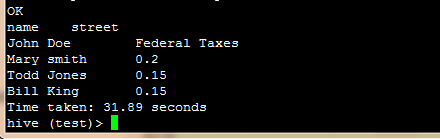
引用map元素，需要指定元素的key

* + 1. hive (test)> select name,deductions["Todd JOnes"] from emp;



Null就是根据键’Todd JOnes’查的，只不过还没有值。

* + 1. hive (test)> select name,address.street from emp;



不管数据的实际意义，只要能按要求查出来即可。

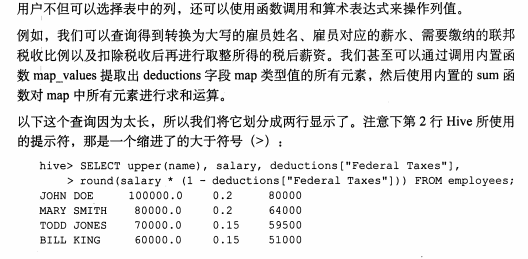
在where子句中同样可以使用这些引用方式。

* 1. 使用正则表达式指定列

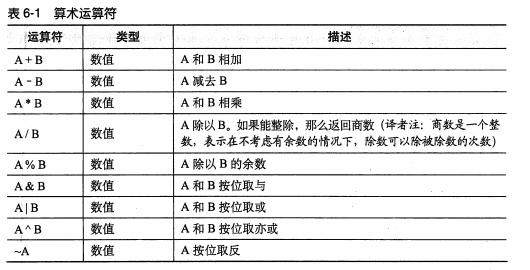
hive (test)> select name,'s.\*' from emp; (0.9版本未实验成功)

注意含有正则的段用单引号括起来，选出所有以s为前缀的列

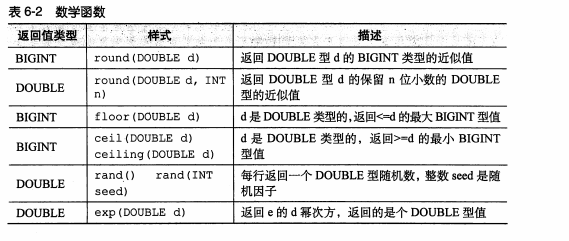
* 1. 使用列值进行计算



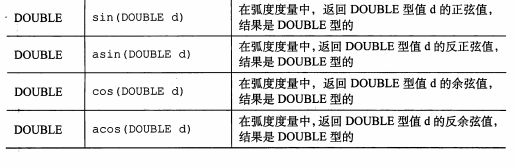
* 1. 算术运算符

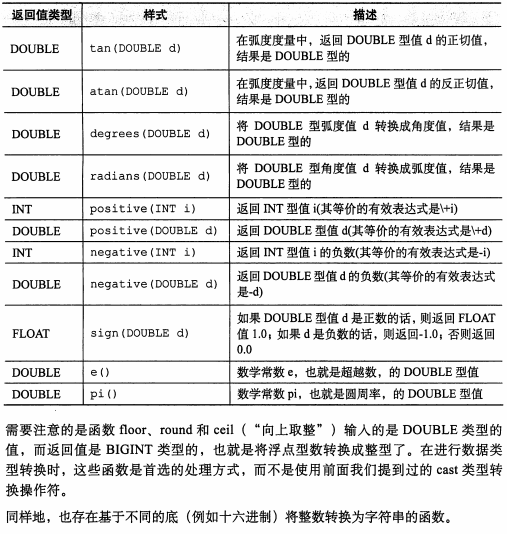


* 1. 使用函数（hadoop的内置函数）
     1. 内置数学函数



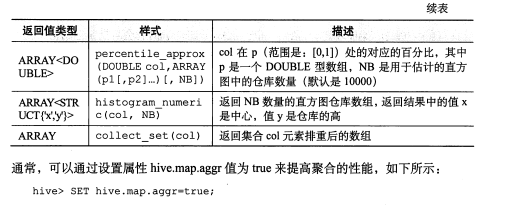




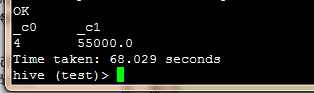


* + 1. 聚合函数

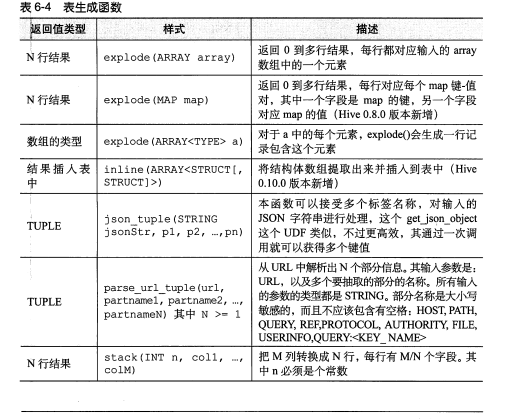




hive (test)> select count(\*),avg(salary) from emp;

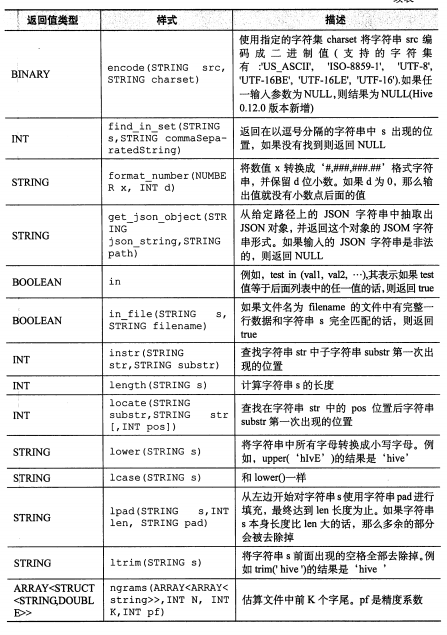


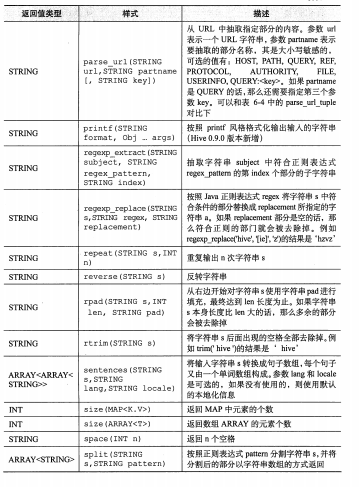
* + 1. 表生成函数(与聚合函数”相反”，将单列扩展为多列或者多行)

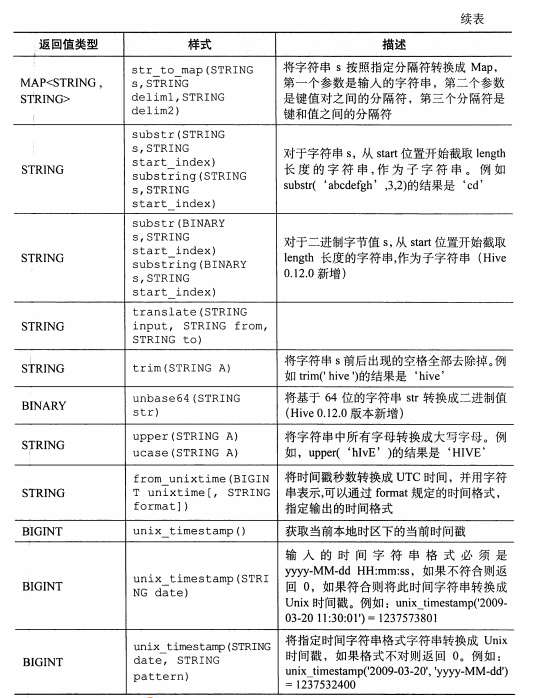


* + 1. 其他内置函数







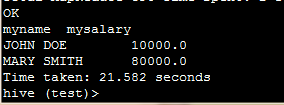




* + 1. LIMIT语句(用于显示返回的行数，与mysqllimit类似)

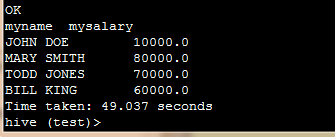
hive (test)> select upper(name) as myName,salary as mySalary from emp limit 2;

只显示查询出的数据的前两条



* + 1. 列别名(就是给复杂的列表达式取一个简单的名称，简化操作，如同sql语句中的别名)

Hive> select upper(name) as myName,salary as mySalary from emp;



* + 1. 嵌套select语句(与sql语法中的嵌套是一样的用法，不再赘述)
    2. CASE…WHEN…THEN句式(和if条件语句类似，处理单个列的查询结果。将then后的语句，作为新的列输出)

hive (test)> select upper(name) name,salary,

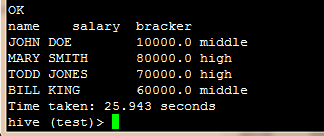
CASE WHEN salary <5000.0 then 'low'

WHEN salary >=5000.0 and salary <70000.0 then 'middle'

WHEN salary >= 70000.0 and salary <90000.0 then 'high'

ELSE 'very high'

END as bracker from emp;



* + 1. Hive可以避免MapReduce的情况

直接查询的模式(本地模式):

Select \* from emp; 无条件查询

Select \* from emp where country =’US’ and state =’CA’

Limit 100; 根据分区字段查询

如何没有设置属性hvie.exec.mode.local.auto设置为true,

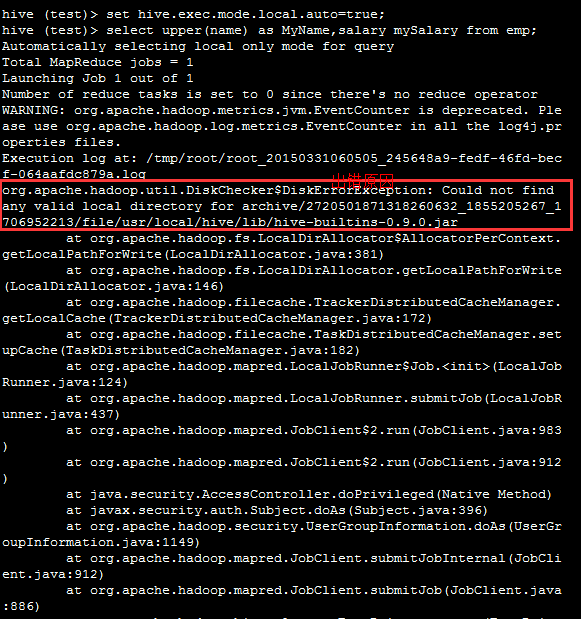
set hvie.exec.mode.local.auto=true;

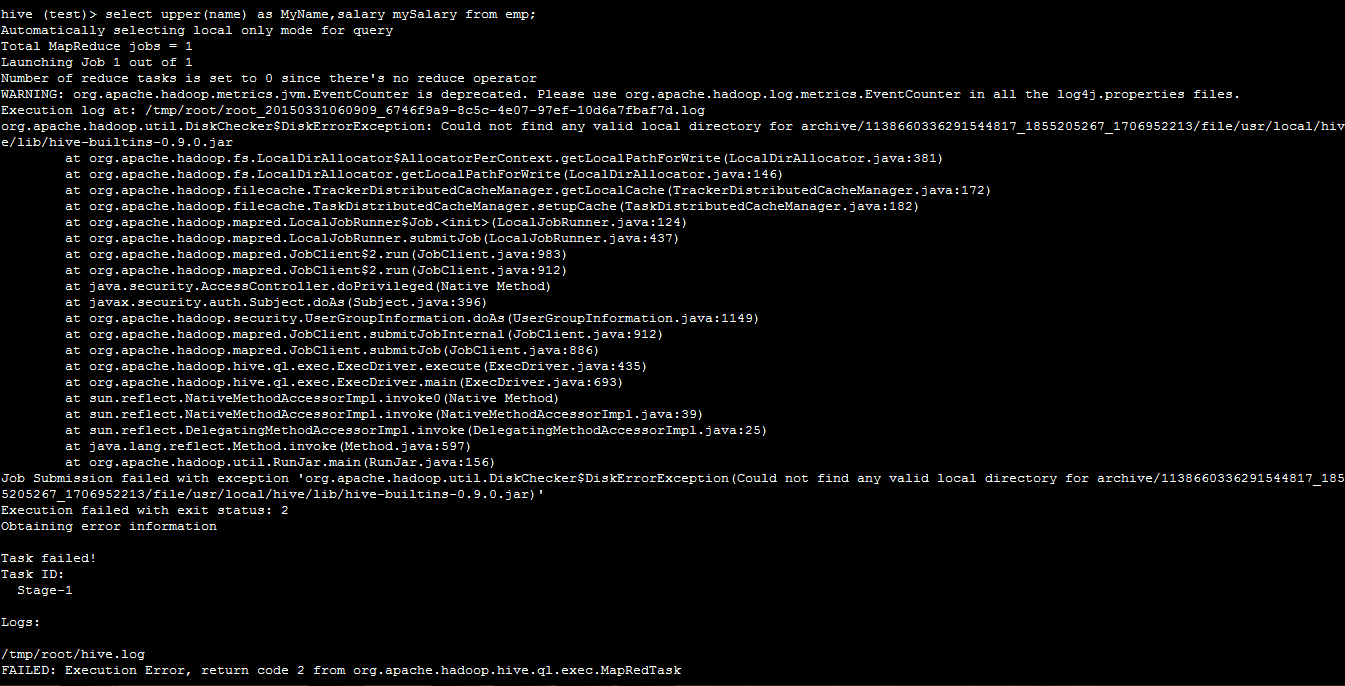
其他查询使用MapReduce执行查询

指南上提示:最好将set hvie.exec.mode.local.auto=true;

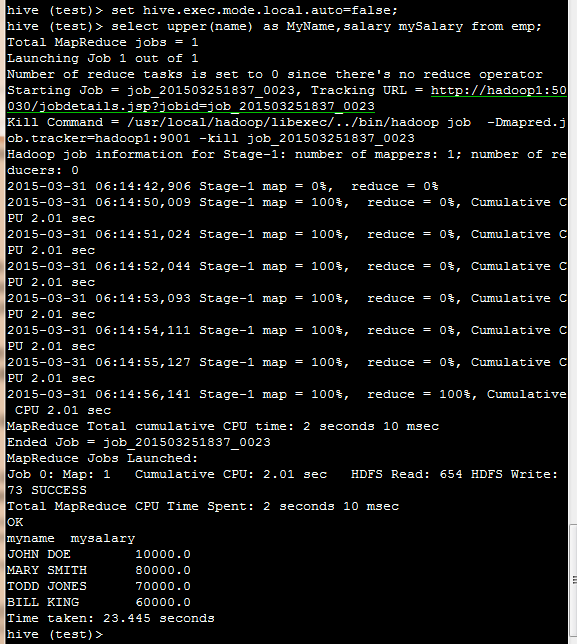
设置添加到$HOME/.hiverc中

但是我在自己搭建的集群环境中确是执行错误，未深入研究。只把错误代码贴出来





将自动本地化执行的属性设置为false后，执行就正确了，初步判断该设置在集群环境中最好设置为false；



* 1. WHERE语句

Where语句使用谓词表达式，可以使用and或者or连接，当查询符合where表达式时，相应的行将被保留输出。

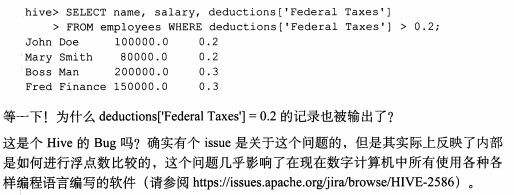
* + 1. 谓词操作符

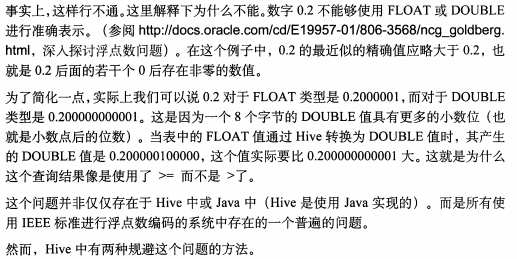


* + 1. 浮点数比较

在编程指南中，hive会将浮点数保存为double类型，在浮点数值比较时，如果没有指定等于，只是大于也会将等于的内容显示出来。但在实际的操作中该问题是不存在的，初步推断应该是版本问题。

Hive编程指南中的例子：



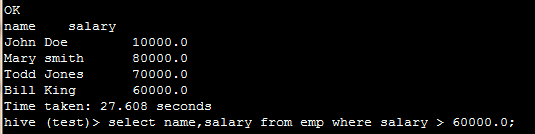


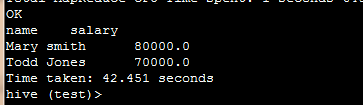
这就是程序上常见的float和double存值时不是准确的值，而是根据不同精度的近似值，这样的比较在编写程序时，也是需要注意的。

在java中可以这样写0.4f代表浮点数，但是在hive中不支持。因此在处理浮点数时要极度谨慎。

但实际执行的情况(未出现该情况，或许是该版本0.9.0已经做了修复)：

hive (test)> select name,salary from emp;





* + 1. Like和RLike

Like的用法和sql的用法相同，不再赘述.

RLike 是hive的功能扩展，可以通过java正则表达式来指定匹配条件，

RLike ‘正则表达式’即可。正则表达式的用法，本例不做介绍。

* 1. GROUP BY语句(用法和sql相同)

HAVING 语句 (用法和sql相同)

* 1. JOIN 语句，支持通常的sql join语句，但是只支持等值连接(连接条件只能是=，不能是>或者<.因为在MapReduce中很难实现这种连接，但是在pig中可以)。

PS：hive中不支持on子句中使用or

* + 1. Inner join（内连接）意义与sql的相同

Hive对每一join连接对象启动一个MapReduce任务，如果有三张表a,b,c连接，则a和b会启动一个MapReduce，然后将第一个MapReduce的输出和c进行连接操作，启动一个MapReduce。

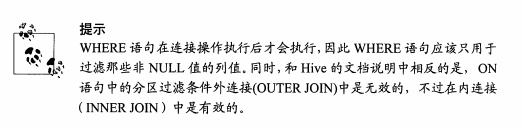
* + 1. JOIN优化

指南上内容未读明白，不知如何总结，暂时为空。P103

* + 1. Left outer join (左外连接)

左表符合where子句的所有记录返回，右表指定选择列的值如果有返回值，无则返回NULL。

* + 1. Outer join(外连接)对于外连接会忽略掉分区过滤条件。



* + 1. Right outer join 返回右表符合where语句的记录，左表匹配不上的用NULL代替。
    2. Full outer join (满外连接)就是左外连接和右外连接的结合。

在oracle中有这种连接方式。

* + 1. Left semi-join (左半开连接)

就是在内连接的基础上只显示左表的指定字段。

PS：hive中不支持右半开连接。

* + 1. 笛卡尔积join(左边表的行数乘右表的行数等于笛卡尔积结果

集的大小)。用错了连接语法可能导致执行很长时间、运行缓慢的笛卡尔积查询。也就是没有添加on子句。

PS:属性hive.mapred.mode值设置为strict时，hive会阻止执行笛卡尔积查询。

* + 1. Map-side join
  1. Order by 和sort by

Order by与sql语法语义相同，是对结果全部排序。所有的数据最后都会通过一个reducer处理过程，对于大数据集，执行时间会很漫长。

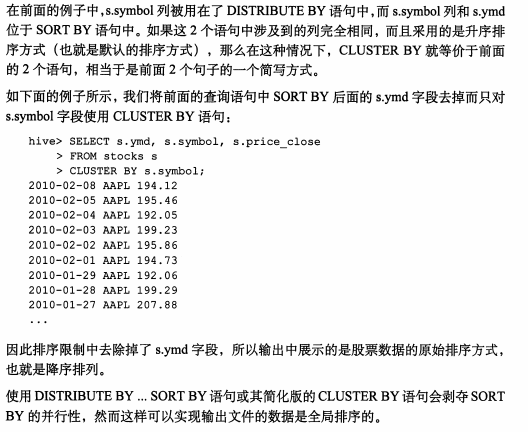
Sort by 是对每个reducer 中对数据进行局部排序，然后进行全局排序，提高全局排序的效率。

如果属性hive.mapred.mode的值是strict，hive会要求使用order by操作必须加上limit语句限制。默认是nonstrict，也就不会有这样的限制。

* 1. 含有sort by 的distribute by

distribute by 控制map输出在reducer中是如何划分的。

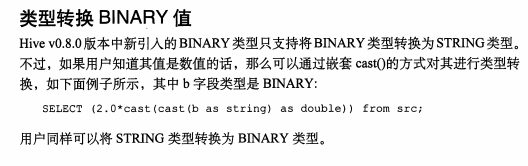
* 1. Cluster by



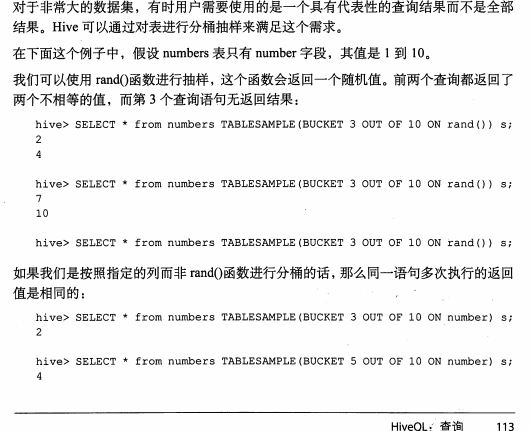
* 1. 类型转换

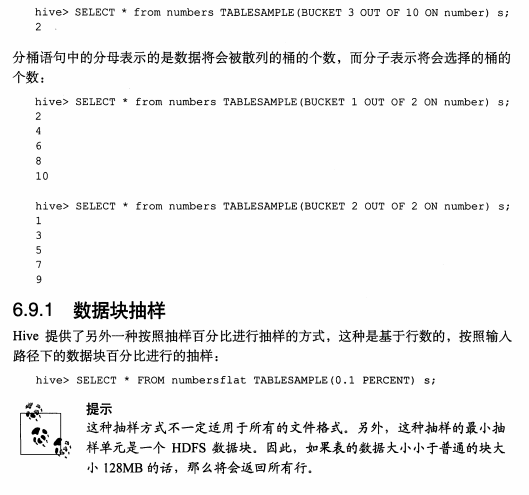
Cast(salary as float) 就将salary字段的类型转换为float类型。

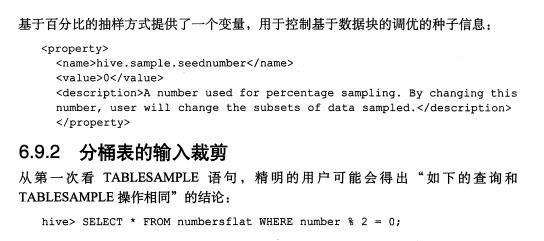


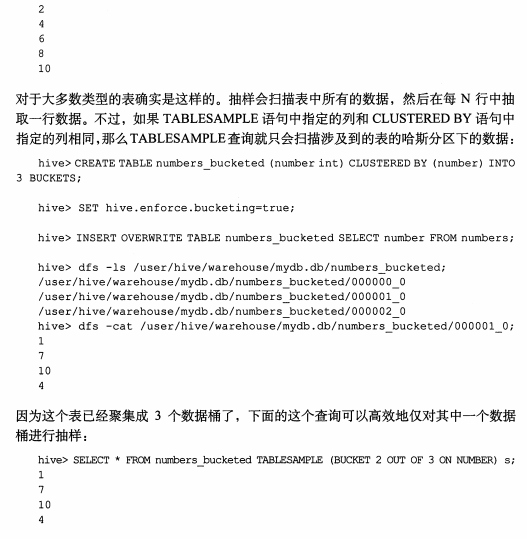


* 1. 抽样查询









* 1. Union all

