

密级：机密(版权所有，翻版必究)



分册名称：

第 册/共 册

大数据高端人才培养计划

Hive 空值处理



沈阳昊宸科技有限公司

2017 年 3 月 18 日

变更履历

修改编号	版本	修改内容	修改人	修改日期
			dulm	



目 录

1 Hive 空值问题	4
-------------------	---



1 Hive 空值问题

Hive 的使用中不可避免的需要对 null、'' (空字符串) 进行判断识别。但是 hive 有别于传统的数据库。

(1) 不同数据类型对空值的存储规则。

int 与 string 类型数据存储, null 默认存储为 \N。

string 类型的数据如果为 "", 存储则是 ""。

另外往 int 类型的字段插入数据 "" 时, 结果还是 \N。

(2) 不同数据类型, 空值的查询。

对于 int 可以使用 is null 来判断空;

而对于 string 类型, 条件 is null 查出来的是 \N 的数据; 而条件 = '' , 查询出来的是 "" 的数据。

例如:

select name,worklocation[3] from person;

```
hive> select name,worklocation[3] from person;
OK
zhangsan      hangzhou
lisi          NULL
Time taken: 0.158 seconds, Fetched: 2 row(s)
```

select * from person where worklocation[3] is null;

```
hive> select * from person where worklocation[3] is null;
OK
lisi      ["changchun","chengdu","wuhan"]
Time taken: 0.152 seconds, Fetched: 1 row(s)
```

向 person.txt 中添加一条数据。

hadoop fs -appendToFile - /apps/hive/warehouse/hivetest.db/person/person.txt

```
[hive@idh104 ~]$ hadoop fs -appendToFile - /apps/hive/warehouse/hivetest.db/person/person.txt
wangsu shenyang, changchun
AC[hive@idh104 ~]$
```

```
hive> select * from person;
OK
zhangsan      ["beijing","shanghai","tianjin","hangzhou"]
lisi          ["changchun","chengdu","wuhan"]
wangsu        ["shenyang","", "changchun"]
Time taken: 0.178 seconds, Fetched: 3 row(s)
hive>
```

分别用 is null 和 = '' 查询。

```
hive> select * from person where worklocation[1] is null;  
OK  
Time taken: 0.144 seconds  
hive> select * from person where worklocation[1] = '';  
OK  
wangsu ["shenyang","", "changchun"]  
Time taken: 0.145 seconds, Fetched: 1 row(s)
```

所以，判断空时要根据实际的存储来进行判断。在开发过程中如果需要对空进行判断，一定得知道存储的是哪种数据。有个处理空的小技巧，Hive 给出一种并非完美的解决方法——自定义底层用什么字符来表示 NULL。

```
ALTER TABLE b SET SERDEPROPERTIES ('serialization.null.format'='');  
ROW FORMAT DELIMITED NULL DEFINED AS '';
```

这句话的意思是让 null 和""等价，也就是让 null 不显示，因为 null 对开发来说不好操作，可能不同地方代表意义不同，而且转码可能也会有问题，所有用""代替。

```
hive> alter table person set serdeproperties('serialization.null.format'='');  
OK  
Time taken: 0.448 seconds  
hive> select * from person where worklocation[1] = '';  
OK  
Time taken: 0.173 seconds  
hive> select * from person where worklocation[1] is null;  
OK  
wangsu ["shenyang",null, "changchun"]  
Time taken: 0.17 seconds, Fetched: 1 row(s)
```

