Hive常用字符串函数

Hive内部提供了很多操作字符串的相关函数，本文将对其中部分常用的函数进行介绍。

下表为Hive内置的字符串函数，具体的用法可以参见本文的下半部分。

返回类型	函数名	描述	
int	ascii(string str)	返回str第一个字符串的数值	
string	base64(binary bin)	将二进制参数转换为base64字符串	
string	concat(string|binary A, string|binary B...)	返回将A和B按顺序连接在一起的字符串，如：concat('foo', 'bar') 返回'foobar'	
array<struct<string,double>>	context_ngrams(array<array<string>>, array<string>, int K, int pf)	从一组标记化的句子中返回前k个文本	
string	concat_ws(string SEP, string A, string B...)	类似concat() ，但使用自定义的分隔符SEP	
string	concat_ws(string SEP, array<string>)	类似concat_ws() ，但参数为字符串数组	
string	decode(binary bin, string charset)	使用指定的字符集将第一个参数解码为字符串，如果任何一个参数为null，返回null。可选字符集为： 'US_ASCII', 'ISO-8859-1', 'UTF-8', 'UTF-16BE', 'UTF-16LE', 'UTF-16'	
binary	encode(string src, string charset)	使用指定的字符集将第一个参数编码为binary，如果任一参数为null，返回null	
int	find_in_set(string str, string strList)	返回str在strList中第一次出现的位置，strList为用逗号分隔的字符串，如果str包含逗号则返回0，若任何参数为null，返回null。如：find_in_set('ab', 'abc,b,ab,c,def') 返回3	
string	format_number(number x, int d)	将数字x格式化为'#,###,###.##'，四舍五入为d位小数位，将结果做为字符串返回。如果d=0，结果不包含小数点或小数部分	
string	get_json_object(string json_string, string path)	从基于json path的json字符串中提取json对象，返回json对象的json字符串，如果输入的json字符串无效返回null。Json 路径只能有数字、字母和下划线，不允许大写和其它特殊字符	
boolean	in_file(string str, string filename)	如果str在filename中以正行的方式出现，返回true	
int	instr(string str, string substr)	返回substr在str中第一次出现的位置。若任何参数为null返回null，若substr不在str中返回0。Str中第一个字符的位置为1	
int	length(string A)	返回A的长度	
int	locate(string substr, string str[, int pos])	返回substr在str的位置pos后第一次出现的位置	
string	lower(string A) lcase(string A)	返回字符串的小写形式	
string	lpad(string str, int len, string pad)	将str左侧用字符串pad填充，长度为len	
string	ltrim(string A)	去掉字符串A左侧的空格，如：ltrim(' foobar ')的结果为'foobar '	

返回类型	函数名	描述
array<struct<string,double>>	ngrams(array<array<string>>, int N, int K, int pf)	从一组标记化的Returns the top-k句子中返回前K个N-grams
string	parse_url(string urlString, string partToExtract [, string keyToExtract])	返回给定URL的指定部分，partToExtract的有效值包括HOST，PATH， QUERY， REF，PROTOCOL，AUTHORITY，FILE和USERINFO。例如： parse_url('http://facebook.com/path1/p.php?k1=v1&k2=v2#Ref1', 'HOST') 返回 'facebook.com'.。当第二个参数为QUERY时，可以使用第三个参数提取特定参数的值，例如： parse_url('http://facebook.com/path1/p.php?k1=v1&k2=v2#Ref1', 'QUERY', 'k1') 返回'v1'
string	printf(String format, Obj... args)	将输入参数进行格式化输出
string	regexp_extract(string subject, string pattern, int index)	使用pattern从给定字符串中提取字符串。如： regexp_extract('foothebar', 'foo(.*?)(bar)', 2) 返回'bar' 有时需要使用预定义的字符类：使用'₩s' 做为第二个参数将匹配s，'s'匹配空格等。参数index是Java正则匹配器方法group()方法中的索引
string	regexp_replace(string INITIAL_STRING, string PATTERN, string REPLACEMENT)	使用REPLACEMENT替换字符串INITIAL_STRING中匹配PATTERN的子串，例如：regexp_replace("foobar", "oo|ar", "") 返回'fb'
string	repeat(string str, int n)	将str重复n次
string	reverse(string A)	将字符串A翻转
string	rpad(string str, int len, string pad)	在str的右侧使用pad填充至长度len
string	rtrim(string A)	去掉字符串A右侧的空格，如： rtrim(' foobar ') 返回 ' foobar'
array<array<string>>	sentences(string str, string lang, string locale)	将自然语言文本处理为单词和句子，每个句子在适当的边界分割，返回单词的数组。参数lang和local为可选参数，例如：sentences('Hello there! How are you?') 返回(("Hello", "there"), ("How", "are", "you"))
string	space(int n)	返回n个空格的字符串
array	split(string str, string pat)	用pat分割字符串str，pat为正则表达式
map<string,string>	str_to_map(text[, delimiter1, delimiter2])	使用两个分隔符将文本分割为键值对。第一个分隔符将文本分割为K-V对，第二个分隔符分隔每个K-V对。默认第一个分隔符为“，”，第二个分隔符为=
string	substr(string|binary A, int start) substring(string|binary A, int start)	返回A从位置start直到结尾的子串
string	substr(string|binary A, int start, int len) substring(string|binary A, int start, int len)	返回A中从位置start开始，长度为len的子串，如： substr('foobar', 4, 1) 返回 'b'
string	translate(string input, string from, string to)	将input中出现在from中的字符替换为to中的字符串，如果任何参数为null，结果为null
string	trim(string A)	去掉字符串A两端的空格
binary	unbase64(string str)	将base64字符串转换为二进制

返回类型	函数名	描述
string	upper(string A) ucase(string A)	返回字符串A的大写形式

1.、字符串长度计算函数：length

语法: length(string A)，
返回值: int
说明：返回字符串A的长度
例子：

hive> select length('iteblog') from iteblog;
7

2.、字符串反转函数：reverse

语法: reverse(string A)
返回值: string
说明：返回字符串A的反转结果
例子：

hive> select reverse(iteblog') from iteblog;
golbeti

3.、字符串连接函数：concat

语法: concat(string A, string B...)
返回值: string
说明：返回输入字符串连接后的结果，支持任意个输入字符串
例子：

hive> select concat('www','.iteblog','.com') from iteblog;
www.iteblog.com

4、带分隔符字符串连接函数：concat_ws

语法: concat_ws(string SEP, string A, string B...)
返回值: string

说明：返回输入字符串连接后的结果，SEP表示各个字符串间的分隔符
例子：

```
hive> select concat_ws('.','www','iteblog','com') from iteblog;
www.iteblog.com
```

5、字符串截取函数：substr,substring

语法: substr(string A, int start),substring(string A, int start)
返回值: string
说明：返回字符串A从start位置到结尾的字符串
例子：

```
hive> select substr('iteblog',3) from iteblog;
eblog
hive>  selectsubstr('iteblog',-1) from iteblog;
g
```

6、 字符串截取函数：substr,substring

语法: substr(string A, int start, int len),substring(string A, intstart, int len)
返回值: string
说明：返回字符串A从start位置开始，长度为len的字符串
例子：

```
hive> select substr('abcde',3,2) from iteblog;
cd
hive> select substring('abcde',3,2) from iteblog;
cd
hive>select substring('abcde',-2,2) from iteblog;
de
```

7、字符串转大写函数：upper,ucase

语法: upper(string A) ucase(string A)
返回值: string

说明：返回字符串A的大写格式
例子：

```
hive> select upper('abSEd') from iteblog;
ABSED
hive> select ucase('abSEd') from iteblog;
ABSED
```

8、字符串转小写函数：lower,lcase

语法: lower(string A) lcase(string A)
返回值: string
说明：返回字符串A的小写格式
例子：

```
hive> select lower('abSEd') from iteblog;
absed
hive> select lcase('abSEd') from iteblog;
absed
```

9、去空格函数：trim

语法: trim(string A)
返回值: string
说明：去除字符串两边的空格
例子：

```
hive> select trim(' abc ') from iteblog;
abc
```

10、左边去空格函数：ltrim

语法: ltrim(string A)
返回值: string
说明：去除字符串左边的空格
例子：

```
hive> select ltrim(' abc ') from iteblog;
abc
```

11、右边去空格函数：rtrim

语法: rtrim(string A)
返回值: string
说明：去除字符串右边的空格
例子：

```
hive> select rtrim(' abc ') from iteblog;
abc
```

12、正则表达式替换函数：regexp_replace

语法: regexp_replace(string A, string B, string C)
返回值: string
说明：将字符串A中的符合java正则表达式B的部分替换为C。注意，在有些情况下要使用转义字符,类似oracle中的regexp_replace函数。
例子：

```
hive> select regexp_replace('foobar', 'oo|ar', '') from iteblog;
fb
```

13、正则表达式解析函数：regexp_extract

语法: regexp_extract(string subject, string pattern, int index)
返回值: string
说明：将字符串subject按照pattern正则表达式的规则拆分，返回index指定的字符。
例子：

```
hive> select regexp_extract('foothebar', 'foo(.*?)(bar)', 1) fromiteblog;
the
hive> select regexp_extract('foothebar', 'foo(.*?)(bar)', 2) fromiteblog;
bar
hive> select regexp_extract('foothebar', 'foo(.*?)(bar)', 0) fromiteblog;
foothebar
```

注意，在有些情况下要使用转义字符，下面的等号要用双竖线转义，这是java正则表达式的规则。

```
select data_field,
    regexp_extract(data_field,'.*?bgStart₩₩=([^&]+)',1) as aaa,
    regexp_extract(data_field,'.*?contentLoaded_headStart₩₩=([^&]+)',1) as bbb,
    regexp_extract(data_field,'.*?AppLoad2Req₩₩=([^&]+)',1) as ccc
    from pt_nginx_loginlog_st
    where pt = '2012-03-26'limit 2;
```

14、URL解析函数：parse_url

语法: parse_url(string urlString, string partToExtract [, stringkeyToExtract])
返回值: string
说明：返回URL中指定的部分。partToExtract的有效值为：HOST, PATH, QUERY, REF, PROTOCOL, AUTHORITY, FILE, and USERINFO.
例子：

```
hive> select parse_url('http://iteblog.com?weixin=iteblog_hadoop', 'HOST') from iteblog;
iteblog.com
hive> select parse_url('http://iteblog.com?weixin=iteblog_hadoop',
    > 'QUERY','weixin') from iteblog;
iteblog_hadoop
```

15、json解析函数：get_json_object

语法: get_json_object(string json_string, string path)
返回值: string
说明：解析json的字符串json_string,返回path指定的内容。如果输入的json字符串无效，那么返回NULL。
例子：

```
hive> select get_json_object('{"store":
>  {"fruit":₩[{"weight":8,"type":"apple"},{"weight":9,"type":"pear"}],
>   "bicycle":{"price":19.95,"color":"red"}
>  },
```

```
> "email":"amy@only_for_json_udf_test.net",
>  "owner":"amy"
> }
> ','$.owner') from iteblog;
amy
```

16、空格字符串函数：space

语法: space(int n)
返回值: string
说明：返回长度为n的字符串
例子：

```
hive> select space(10) from iteblog;
hive> select length(space(10)) from iteblog;
10
```

17、重复字符串函数：repeat

语法: repeat(string str, int n)
返回值: string
说明：返回重复n次后的str字符串
例子：

```
hive> select repeat('abc',5) from iteblog;
abcabcabcabcabc
```

18、首字符ascii函数：ascii

语法: ascii(string str)
返回值: int
说明：返回字符串str第一个字符的ascii码
例子：

```
hive> select ascii('abcde') from iteblog;
97
```

19、左补足函数：lpad

语法: lpad(string str, int len, string pad)
返回值: string
说明：将str进行用pad进行左补足到len位
例子：

hive> select lpad('abc',10,'td') from iteblog;
tdtdtdtabc

注意：与GP，ORACLE不同，pad 不能默认

20、右补足函数：rpad

语法: rpad(string str, int len, string pad)
返回值: string
说明：将str进行用pad进行右补足到len位
例子：

hive> select rpad('abc',10,'td') from iteblog;
abctdtdtdt

21、分割字符串函数: split

语法: split(string str, stringpat)
返回值: array
说明: 按照pat字符串分割str，会返回分割后的字符串数组
例子：

hive> select split('abtcdtef','t') from iteblog;
["ab","cd","ef"]

22、集合查找函数:find_in_set

语法: find_in_set(string str, string strList)

返回值: int
说明: 返回str在strlist第一次出现的位置，strlist是用逗号分割的字符串。如果没有找该str字符，则返回0
例子：

```
hive> select find_in_set('ab','ef,ab,de') from iteblog;
2
hive> select find_in_set('at','ef,ab,de') from iteblog;
0
```