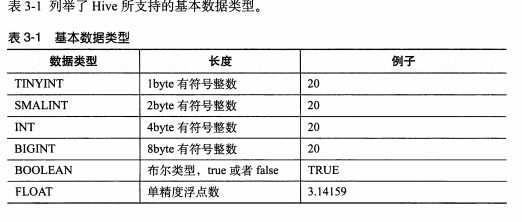
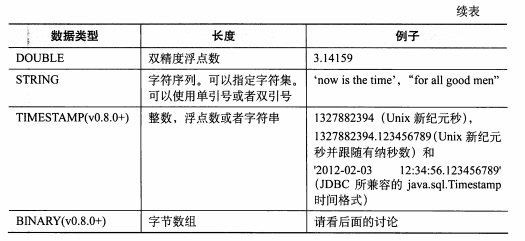
Hadoop学习之Hive编程指南学习(二)

1. 基本数据类型

Hive支持关系型数据库中的大多数基本数据类型，也执行关系DB中很少出现那的三种集合数据类型。





Hive通常不限制列的长度，强调优化磁盘的读写性能。而sql方言中提供限制最大长度的”字符数组”类型在hive中不支持，sql中是为了性能优化，容易建立所以，数据扫描。

新增数据类型TIMESTAMP的值可以是整数，也可以是unix时间秒数，也可以是浮点数，精确到纳秒；还可以是字符串，JDBC所约定的时间字符串格式(YYYY-MM-DD hh:mm:ss.fffffffff)。

TIMESTAMP表示的是UTC时间，可以再不同时区间相互转换的内置函数to\_utc\_timestamp和from\_utc\_timestamp函数。

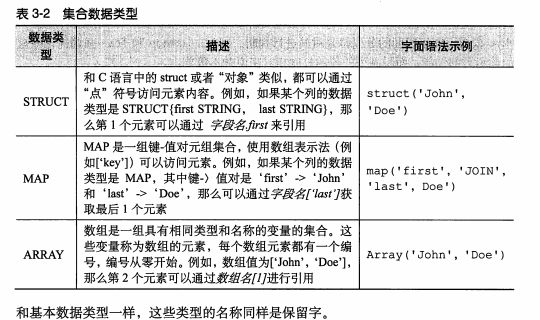
BINARY数据类型和关系DB中的VARBINARY类型类似，但和BLOB类型不同，因为BINARY的列是存储在记录中的，而BLOB不同，BINARY可以记录中包含任意字节，这样可以防止hive尝试将其作为数字，字符串等解析。

如果将float类型的列和double类型的列做对比或者将整型和另一种整型的值作对比，结果就是hive会隐式的将两种类型转换为比较大的一种类型。

如果希望将字符串列转为数字，可以显式转换，cast(s AS INT)

1. 集合数据类型

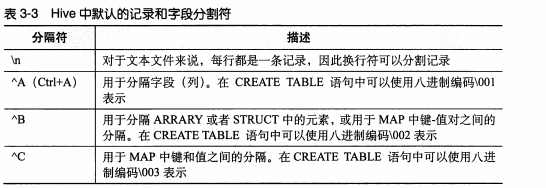
Hive中的列支持使用struct，map和array集合数据类型。



在传统DB中没有这三种类型是因为它们会趋向于破坏标准格式，增大数据冗余的风险，但在大数据系统中，好处就是可以提供更高吞吐量的数据。

PS：Hive中并没有键的概念，但用户可以对表建立索引。

1. 文本文件数据编码



事实上，hvie目前为止对于linesterminated by .. 仅支持’\n’,也就是说行与行之间的分合富只能是’\n’，因此这个子句现在使用起来还是有限制的。

虽然用户可以明确指定这些子句，但在大多数情况下，大多子句还是使用默认的分隔符的，只需要明确指定哪些需要替换的分隔符即可，用户可以使用’,’作为分隔符，也可以使用’\t’作为字段分隔符。

这种强大的可定制功能使得可以很容易的使用hive处理由其他工具和各种各样的ETL（数据抽取、数据转换、数据装载）程序产生的文件。

1. 读取模式

当用户向传统DB写数据时，不管是采用装在外部数据，还是讲查询的输出结果写入，或者update语句，数据库对存储具有完全的控制力。数据库就是守门人。传统DB是写时模式，即数据在写入数据库时对模式进行检查。

Hive对底层存储并没有这样控制，因此Hive不会在数据加载时验证，而是在查询时进行，也就是读时模式。

那么如果模式和文件内容并不匹配将会怎样呢？hive对此做的非常好，因为可以读取这些数据。如果每行记录中的额字段个数少于对应的模式中定义的字段个数的话，那么用户将会看到查询结果中有很多的null值。如果那些字段是数值型的，但是hive在读取时发现存在非数值行的字符串值得话，那么对于那血字段将会返回null值。除此之外的其他情况下，hive都机理尝试尽可能的将各种错误恢复过来。