Character String Data Types

Hive and Impala support three character data types: STRING, CHAR, and VARCHAR. These types represent alphanumeric text values.

The STRING data type represents a sequence of characters with no specified length constraint.*

If you're familiar with relational databases, you are probably more accustomed to the character types CHAR and VARCHAR, which have a specified length.

The CHAR type represents fixed-length character sequences, with a precise specified length. Values longer than the specified length are truncated. Values shorter than the specified length are padded with spaces. If you assign the 13-character value Impala rules! to a CHAR column with length 16, then Hive and Impala will pad that value with three spaces to make it 16 characters long: Impala rules!␣␣␣ (The three symbols shown in this example represent spaces.)

The VARCHAR type represents character sequences with a maximum specified length.

Values longer than the maximum are truncated, but values shorter than the maximum are not padded with spaces. If you attempt to assign the 13-character value Impala rules! in a VARCHAR column with a maximum length of 10, then Hive and Impala will truncate that value to 10 characters, discarding the last three characters: Impala rul. However, if the maximum length is 13 or more, the stored value will be exactly Impala rules! (with no extra spaces as you would get with the CHAR type).

The table here summarizes these examples.

Data Type	Description	Value (attempting Impala rules!)

STRING	Any number of characters	Impala rules!
CHAR(10)	Exactly 10 characters	Impala rul
CHAR(16)	Exactly 16 characters	Impala rules! ␣ ␣ ␣
VARCHAR(10)	At most 10 characters	Impala rul
VARCHAR(16)	At most 16 characters	Impala rules!

With CHAR types, trailing spaces are ignored in comparisons. With VARCHAR and STRING values, any trailing spaces are considered in comparisons. (This makes sense, since neither is automatically padded—trailing spaces are not considered to be “padding” in these cases.)

You should generally choose STRING over CHAR or VARCHAR. STRING offers greater flexibility and ease of use, and in some cases Hive and Impala have better performance and compatibility when using STRING columns. But if you have a particular need for string values with precise lengths or with maximum lengths, then you could use CHAR or VARCHAR.

*Footnote: Actual String Limits

There actually are practical limits to the length of strings, though in most real-world applications, it's unlikely you'll ever come up against them. For example, in Impala, these are the considerations for lengths of strings (taken from STRING Data Type in Cloudera's Impala documentation):

- The hard limit on the size of a STRING and the total size of a row is 2GB.

- If a query tries to process or create a string larger than this limit, it will return an error to the user.
- The limit is 1GB on STRING when writing to Parquet files.
- Queries operating on strings with 32KB or less will work reliably and will not hit significant performance or memory problems (unless you have very complex queries, very many columns, and so on.)
- Performance and memory consumption may degrade with strings larger than 32KB.

This varies somewhat according to which version of Impala you are using, so if you are working with exceptionally large strings, check the documentation.