Character Sets

NOTE: Although this reading is in the Cloud Storage lesson, it applies to working with data in HDFS as well.

As you might know, computer systems use defined *character sets,* which are collections of characters, for what are allowable characters. Examples include Unicode, ASCII, Extended ASCII, and various ISO sets.

Compatibility between the character sets your data files use and the character set your system uses is an important consideration. Incompatible characters may appear oddly, or they might even cause the system to throw errors on reading.

If you're familiar with SQL on RDBMSs, you might have seen the CHARACTER SETclause in the LOAD DATA statement, which lets you specify the character set when you load the data. Hive and Impala do not have such a clause in their LOAD DATA statements.

Cloudera's documentation for Impala says this about the character sets used by Impala:

> For full support in all Impala subsystems, restrict string values to the ASCII character set. Although some UTF-8 character data can be stored in Impala and retrieved through queries, UTF-8 strings containing non-ASCII characters are not guaranteed to work properly in combination with many SQL aspects.…

> For any national language aspects such as collation order or interpreting extended ASCII variants such as ISO-8859-1 or ISO-8859-2 encodings, Impala does not include such metadata with the table definition. If you need to sort, manipulate, or display data depending on those national language characteristics of string data, use logic on the application side.

Hive does have full support for the UTF-8 character set, but no others. From the Hive User FAQ:

You can use Unicode string on data/comments, but cannot use for database/table/column name.
You can use UTF-8 encoding for Hive data. However, other encodings are not supported (HIVE-7142 introduce encoding for LazySimpleSerDe, however, the implementation is not complete and [does] not address all cases).

Both Hive and Impala also include string functions (such as encode() and decode() in Hive, and base64encode() and base64decode() in Impala) that might help for transmitting data in characters other than UTF-8 or ASCII.