Creating Partitioned Tables

To create a partitioned table, use the PARTITIONED BY clause in the CREATE TABLE statement. The names and types of the partition columns must be specified in the PARTITIONED BY clause, and only in the PARTITIONED BY clause. They must not also appear in the list of all the other columns.

```
CREATE TABLE customers_by_country

    (cust_id STRING, name STRING)

PARTITIONED BY (country STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
```

The example CREATE TABLE statement shown above creates the table customers_by_country, which is partitioned by the STRING column named country. Notice that the country column appears only in the PARTITIONED BY clause, and not in the column list above it. This example specifies only one partition column, but you can specify more than one by using a comma-separated column list in the PARTITIONED BY clause. Aside from these specific differences, this CREATE TABLE statement is the same as the statement used to create an equivalent non-partitioned table.

Table partitioning is implemented in a way that is mostly transparent to a user issuing queries with Hive and Impala. A partition column is what's known as a *virtual column*, because its values are not stored within the data files. Following is the result of the DESCRIBE command on customers_by_country; it displays the partition column country just as if it were a normal column within the table. You can refer to partition columns in any of the usual clauses of a SELECT statement.

name	type	comment

cust_id	string	
name	string	
country	string	

Note: In the previous lesson, you learned about using COMPUTE STATS in Impala and ANALYZE TABLE … COMPUTE STATISTICS in Hive. If the table is partitioned, Impala cannot use Hive-generated column statistics, so for partitioned tables, it's best to compute the statistics with the engine you'll be using to query the table.

Try It!

Use the CREATE TABLE command above to create a customers_by_country table on your VM, then use DESCRIBE to show the columns. Notice that it's just as described above: there is no apparent difference between the partition column and the other columns.

You'll use this table in the next readings too, so don't drop it yet.