SQL LOAD DATA Statements

One way to load data into a table is by running a LOAD DATA INPATH statement with Hive or Impala. This moves the specified data files into the table's storage directory in the file system (HDFS or S3).

The example shown below moves the file sales.txt from the HDFS directory /incoming/etl to the directory for the table named sales. Notice that this statement specifies the destination as a table name, not a directory; Hive or Impala uses metadata from the metastore to determine the table's storage directory, and moves the file there.

```
LOAD DATA INPATH '/incoming/etl/sales.txt' INTO TABLE sales;
```

The source path can refer either to a file, as in this example, or to a directory, in which case Hive or Impala will move all the files within that directory into the table.

The LOAD DATA INPATH statement *adds* the source files to any existing files that are already in the table's directory—that is, it does not remove existing files in the table's directory, and if there are any filename collisions, then it automatically renames the new files so that no existing files are overwritten. In some cases, you may want to delete all existing data files in the table's directory before loading new data files. To do this, use the OVERWRITE keyword, as shown in the example below. This option is useful when you need to do a complete reload of all the data in a table.

```
LOAD DATA INPATH '/incoming/etl/sales.txt' OVERWRITE INTO TABLE sales;
```

The LOAD DATA INPATH statement assumes that the files are already somewhere in a file system that is accessible to your instance of Hive or Impala (like HDFS or S3). If they are not, then you first need to load files from your local filesystem into HDFS or S3, for example by running an hdfs dfs -put command, or by using the Hue File Browser. Also, this method *moves* the files rather than copying them; the files will no longer exist in the source directory after the statement is executed.

This probably sounds a lot like using hdfs dfs -mv—and it is, but there are a couple of advantages to using LOAD DATA INPATH.

One is that, if you're running the LOAD DATA INPATH statement with Impala, the metadata cache is automatically updated. You do not need to execute a REFRESH command; your next query will include the new data.

The other advantage is that since the statement renames any files that are the same as files that already exist within the directory, you don't need to worry about what your files are named. If you use hdfs dfs -mv and there is an existing file with the same name, the command will fail and no changes will be made—you'll have to rename one of the files yourself, first.