Load CSV Project

1. **Project Purpose**

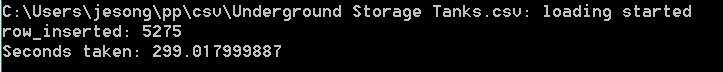
The project aims at make the process of loading source files into Oracle easier for data analysts. It uses a Python script as a vehicle to load source files into Oracle automatically.

1. **What the script can do?**

Currently, the script can only load text based files: CSV or TXT, but it is possible to include Excel file into its capability in the future.

The average loading speed is 1000 rows / minute, if the loading is done from local computer. It could be faster if the script is run on server. Comparison is also done between the script and Oracle sql developer. it takes almost the same time for Oracle sql developer to load the same table from access format into Oracle.

The following screen shot could give you some concept about the speed.



A pressure test was also done on the script. During the test, 100,000 records were inserted. It took around 90 minutes to load these records.

Files could be handled

1. CSV file

( a )CSV file following the most strict rules

For these csv files, their fields are delimited by comma and all its text are wrapped by double quotation mark.

Example:

"9711","2011BAZ6527","123 AUTO DR BOISE"

( b )CSV file with Excel flavor

Usually, they are exported from excel. Not every field is wrapped by double quotation mark. It is only used when really necessary, and double quotation mark within a field is escaped by another double quotation mark.

Example:

19-Jul-16,"5'4"""

( c ) CSV file having CLOB fields with new line and carriage return in it

Example:

"The Boeing (MD-St. Louis) site, formerly the McDonnell Douglas Corp., operated an aerospace manufacturing facility at the Tract I location, located at McDonnell Douglas Boulevard and Banshee Road in Hazelwood. Tract 1 occupies about 228 acres next to Lambert St. Louis International Airport. In 1941, McDonnell Douglas began manufacturing transport aircraft, aircraft components, space systems/missiles and combat aircraft, which included fabricating aluminum, titanium, composite structures and other airframe material; chemical processing; degreasing; painting; and aircraft assembly, fueling and flight testing.

The site activities produced approximately 48 different hazardous waste streams. McDonnell Douglas stored the waste in various storage facilities around the site that were permitted under two hazardous waste permits, one issued by the department and one issued by the Environmental Protection Agency (EPA), both effective March 5, 1997. The department issued the Missouri Hazardous Waste Management Facility Part I Permit. EPA issued the Hazardous and Solid Waste Amendments Part II Permit. McDonnell Douglas also distilled spent solvents under a Resource Recovery Certification, issued by the department.

"

1. Txt file

Most txt files are delimited by a character, such as ‘\t’. Anyway, you need to tell the script what delimiting character is used.

Example:

460504 UST CR Single 1998-01-15 00:00:00.0 1000 6 8

1. **What the script cannot do?**
2. Excel file

The script could not handle Excel file currently.

1. CSV file in Unicode and having ASCII NUL characters

The script uses Python CSV module to parse CSV file. The module does not support Unicode and files containing ASCII NUL characters.

1. The script cannot handle different file types at the same time, so you cannot let the script load txt and csv in one batch.
2. **How to use the script**

The script name is ‘load\_csv.py’. It takes three parameters currently:

Directory in absolute format: it is the directory in which you put all your files to be loaded

Prefix: the string you want to put in front of the table name

Delimiter: it is optional, but if the fields are not delimited by comma, you need to specify

(1) Normal csv file, you write like this:



(2) Normal txt file, you write like this



\t stands for tab

(3) In case you forget what parameters to write, correct format is printed for your reference.



Excel uses a serial number starting from 1899-12-30 to express date. What you see in Excel formula bar actually is only an object formatted according to your operating system default configuration. So if you want to keep that, you still need to do “Text to Column” in Excel.