**Large file processor**

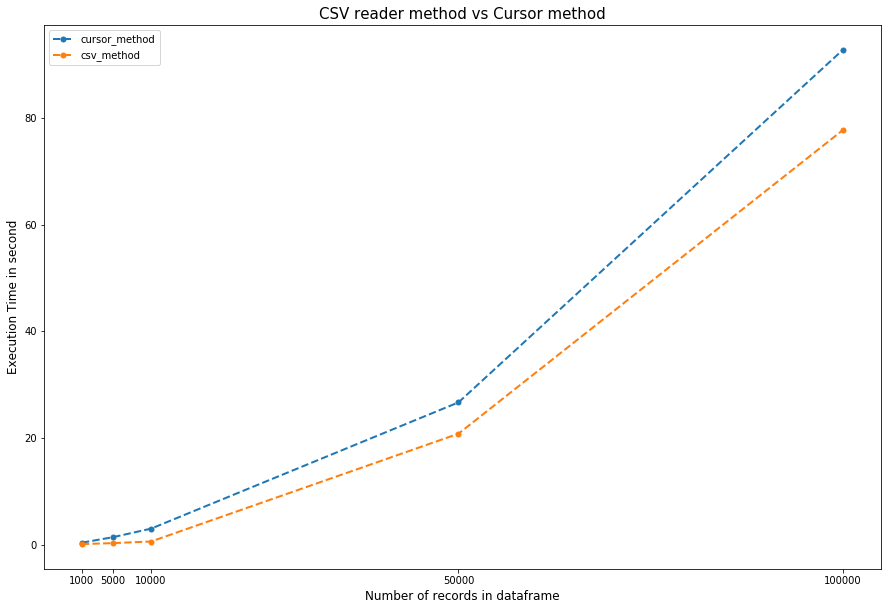
**Problem statement:**

We need to be able to import products from a CSV file and into a database. There are half a million product details to be imported into the database. You can find the CSV file here in a compressed format Large file processing Assignment. After the import, we will run an aggregate query to give us no. of products with the same name.

**Workaround:**

I could be able to solve the problem statement described above. Please find below an explanation of the concept as well as the steps to carry out the program.

* Initially, I tried to predict mysql methods in order to determine the quickest way to load data into the database.
* I contrasted the MySQL Cursor method and the CSV file reader method.
* As shown in the image below, the csv file reader method appears to be efficient, so I used it.



* To improve the output, I updated the duplicate values in the ‘sku’ column using pandas.
* I used the csv file reader method to load the data into a MySQL database.
* After uploading the data, used the aggregate query to get the number of items with the same name in a separate table. I also saved the table for your reference.

In the Jupyter notebook, I have provided a step-by-step explanation as well.

**Deliverable:**

I have uploaded the code in GitHub

GitHub link: <https://github.com/parthiban-jayachandran/Large_File_Processor>

**a. Steps to run your code.**

* Please update the database information in two places, as I also mentioned in the Jupyter notebook.
* Incase if you are going to use the different file then the file name also needed to be updated

**b. Details of all the tables and their schema, [with commands to recreate them]**

mysql> *USE POSTMAN;*

Database changed

mysql> *DESC PRODUCT\_DATA;*

+-------------+--------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+-------------+--------------+------+-----+---------+-------+

| name | varchar(255) | NO | | NULL | |

| sku | varchar(255) | NO | PRI | NULL | |

| description | varchar(255) | NO | | NULL | |

+-------------+--------------+------+-----+---------+-------+

3 rows in set (0.01 sec)

mysql> *DESC PRODUCT\_COUNT;*

+-----------------+--------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+-----------------+--------------+------+-----+---------+-------+

| name | varchar(255) | NO | | NULL | |

| no. of products | bigint | NO | | 0 | |

+-----------------+--------------+------+-----+---------+-------+

2 rows in set (0.00 sec)

**c. What is done from “Points to achieve” and number of entries in all your tables with sample 10 rows from each**

mysql> SELECT \* FROM PRODUCT\_DATA LIMIT 10;

+-----------------------+---------------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+

| name | sku | description |

+-----------------------+---------------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+

| James Oconnor | a-ability-see-gun | According member fine program. Concern single too ahead my. Loss onto which include listen later present.

Election sport quite notice why. Find system writer might. Sing prevent compare black. |

| Jennifer Lambert | a-above-its-focus | Go audience old. Law main federal area myself.

Leave various leave discover consumer hotel. Safe fall up compare plant affect stuff. |

| Christopher Tate | a-act-cut-either | Data tell enter. Because stock along continue follow respond off value. Trial try exactly type simply full. |

| Vicki Barber | a-act-spring-camera | Difference compare society best structure democratic team machine. Administration item light among.

Each when capital condition election miss defense. Left after treat listen law girl. |

| Tina Dunn | a-activity | Discussion itself those stand beat Mr. Any from event.

Training cultural avoid artist quite say figure. Play fill cultural know education arm rate reflect. Me investment pull star. |

| Melissa Thomas | a-add-parent-wish | Thank rise institution vote beautiful board. Although high mission size crime show natural color. |

| Dr. Monica Oliver PhD | a-address-age | Under similar week explain less. Could which look pull ready.

Trip lot speech ago. Benefit ground daughter court last. Sometimes reduce take chair cold executive now. |

| Candace Francis | a-adult-situation | Fine month cold age teach all factor brother. Table edge believe garden recently. Project both claim positive minute country night form. Site personal to able point of growth. |

| Adrian Smith | a-against-form | Television worker entire produce read. Her society middle.

You claim degree this newspaper position. Family there wrong sign smile. |

| Bailey Irwin | a-age-make-picture | Gun necessary knowledge human leave. Thank his drop which go long. Shoulder our yet fact smile.

Speech campaign charge value. Gun approach nothing physical mind. |

+-----------------------+---------------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+

10 rows in set (0.00 sec)

mysql> select \* from product\_count limit 10;

+-----------------------+-----------------+

| name | no. of products |

+-----------------------+-----------------+

| James Oconnor | 5 |

| Jennifer Lambert | 7 |

| Christopher Tate | 5 |

| Vicki Barber | 1 |

| Tina Dunn | 3 |

| Melissa Thomas | 23 |

| Dr. Monica Oliver PhD | 1 |

| Candace Francis | 1 |

| Adrian Smith | 9 |

| Bailey Irwin | 1 |

+-----------------------+-----------------+

10 rows in set (0.00 sec)

**d. What is not done from “Points to achieve”. If not achieved write the possible reasons and current workarounds.**

I attempted to avoid the duplicates in sku column by using a mysql query, but I was unable to do it. So, I used pandas to process the duplicates and concatenated the duplicate values with the customer's name like the below.

| Adrian Smith | a-against-form

| Bailey Irwin | a-against-form 🡪 | Bailey Irwin | a-against-form-| Bailey

**e. What would you improve if given more days**

I might try to load the data into the database as quickly as possible.

If there is something else I need to provide, please let me know.

----

Parthiban Jayachandran

+91 9500255911

[jparthibanbe@gmail.com](mailto:jparthibanbe@gmail.com)