**Step 1: Single-Process Program**

**Objective**: Develop a single program to create totalfile.txt by summing the corresponding lines of hugefile1.txt and hugefile2.txt.

**Python Implementation**:

A screenshot of a computer

Description automatically generated

**Execution**

Unable to execute

A screenshot of a computer screen

Description automatically generated

**Got only 250M records in 10 mins**

A black screen with white text

Description automatically generated

**Code is working as expected**

A screen shot of a computer

Description automatically generated

**Step 2: Split Files and Parallel Processing**

**Objective**: Break the task into smaller chunks, process concurrently, and reduce execution time.

* **Split the Files into Two Halves:**

Use split in Linux to divide the files:

split -n 2 hugefile1.txt hugefile1\_

split -n 2 hugefile2.txt hugefile2\_

This creates hugefile1\_aa, hugefile1\_ab, hugefile2\_aa, hugefile2\_ab.

* **Create Two Programs:**

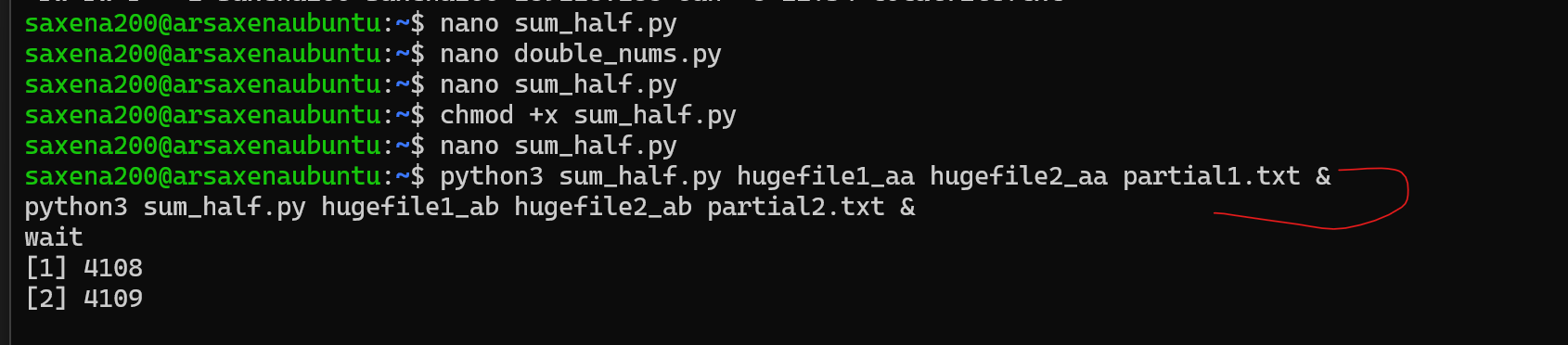
Program 1 processes the first half (hugefile1\_aa and hugefile2\_aa), and Program 2 processes the second half (hugefile1\_ab and hugefile2\_ab). Both write partial results to separate output files.

**Python Code for Each Program**:

A screenshot of a computer

Description automatically generated

* Run the two programs simultaneously:



**CPU Usage**

CPU usage is peaking while I am running the program in parallel

A screenshot of a computer

Description automatically generated

* **Result**

Failed with Space issues

**A screenshot of a computer program

Description automatically generated**

**It wrote around 252M lines before failing**

A screenshot of a computer screen

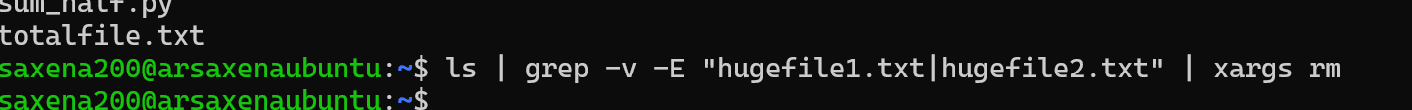
Description automatically generated

**Step 3: Split into 10 Files and Parallelize**

**Objective**: Divide files into 10 smaller chunks and process in parallel.

* Make some space on the system

Delete unused files



* Split Files into 10 Parts:

A black background with white text

Description automatically generated

* **Listing**

A screenshot of a computer screen

Description automatically generated

* **Create a Bash Script:**

Process each pair of split files (hugefile1\_aa and hugefile2\_aa, etc.) in parallel using Python:

#!/bin/bash

for i in {a..j}; do

python3 sum\_half.py hugefile1\_a$i hugefile2\_a$i partial\_$i.txt &

done

wait

cat partial\_\*.txt > totalfile.txt

* **Result**