Assignment 1 - Parallel data analysis using OpenMPI

As part of a data analytics team in a B2B retailer, you have been given a CSV file with sales transaction data from across the globe.

Sales transactions data

You need to create parallel programs that can process this data set efficiently using a cluster of machines. The data contains the following columns:

index, region, country, item_type, sales_channel, order_priority, order_date, order_id, ship_date, units_sold, unit_price, unit_cost, total_revenue, total_cost, total_profit

Here are 2 questions you need to answer from the data :

(a) What is the total **units_sold** in each **region** across the entire data set ?

(b) What is the average **total_profit** per transaction in the entire data set ? Remember that average is not an associative and commutative operation.

Think about how will you partition the data in each of the cases, what computation will happen on each partition of data, how will you "join" all the partial results together and present it on one node ?

You can adopt a master-slave architecture where a specific process is the master that takes the file and performs data partitioning, sends data to slaves, slaves perform parallel computation and send results to the master, master does some last step computation if needed and presents the answer to the user. A master process can also function as a slave. You do not need to adopt any parallel processing technique (e.g. using multi-threading) within the master or the slave to speed up local tasks. Your focus should be on parallel processing across processes.

You need to use OpenMPI to implement the parallel program. Use the standard functions for communication within the cluster of processes. You can run the processes on the same multi-core machine. Given a large data set, you can use the file system to store data partitions, intermediate results at slaves and the final result. Just name the files prefixed with a process identifier so that processes know which data should be read or written by which process. MPI rank can be used as a process identifier.

You can extend this to a cluster of VMs with independent file system storage. But that is **optional learning and not part of this assignment evaluation**. The details on how to run a vanilla MPI cluster without any other cluster manager is given in this blog post below.

https://mpitutorial.com/tutorials/running-an-mpi-cluster-within-a-lan/ (Links to an external site.)

Your assignment output should be a **PDF document** containing the following :

(a) **Design** of your program for each of the 2 cases, e.g. data partitioning, computation at master and slaves etc.

(b) Clear **documented code** in each of the 2 cases

(c) Record the **query results** when you run the code in each of the 2 cases.

(d) Record the end to end **time taken for analysis** in each of the 2 cases with 2, 4, 8 processes. Also specify how many CPU cores you are using for your tests.

General instructions:

- Only one of the students from the group should submit the assignment pdf. DO NOT submit multiple copies from same group.
- Equal contribution from all members in the group will be assumed. So please balance your work within the team.
- Do not follow any unfair practices in copying content from Internet or across groups. Such activities will incur a severe penalty.