EVALUATION OF UNSUPERVISED LEARNING MODELLING WITH PARALLEL PROCESSES

Quynh T Nguyen[1,2], Satyam Vatts[1], Avneet Kaur[1],

Tatjana Jancic-Turner[1], Raouf N.G. Naguib[3]

[1]Mathematics & Statistics Department, Langara College, Vancouver, Canada
[2]Department of Business Administration and Management, Dai Nam University, Hanoi, Vietnam
[3]School of Mathematics, Computer Science Engineering, Liverpool Hope University, Liverpool, UK

UNSUPERVISED MACHINE LEARNING

Clustering Based Approach

CLUSTERING

Technique to form segments of observations based on variations and similarities among them

Heavily utilized to unravel hidden patterns & trends

TECHNIQUES

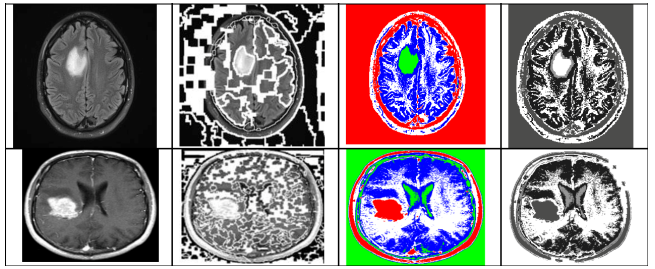
K-Means

K-Means++

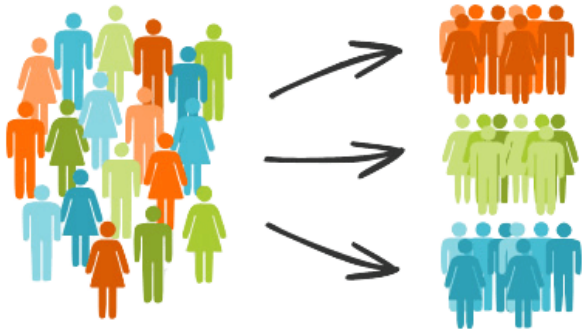
K-Means Parallel

APPLICATIONS

Healthcare: Identifying subgroups of diseases or patients for better diagnosis and treatment



Market Research : Customer Segmentation to discover groups of similar customers



OPTIMIZED TECHNIQUES → SUSTAINABLE FUTURE

Working Towards A Reliable, Optimal & Sustainable Approach

PROBLEM

Clustering is a CPU-Intensive task

MAIN GOAL

To utilize the available compute power optimally for clustering processes

PURPOSE

Aimed at reducing the carbon footprint, less expenditure on infrastructure & making right decision in selecting the hardware

MOTIVATION

Help individual and small enterprises with limited compute infrastructure to develop data-driven products

Help in identifying appropriate approach for expanding or upgrading existing technological infrastructure

EXPERIMENTATION IN DIVERSE ENVIRONMENTS

A Comparative Study To Research A Generic Solution

ON-PREMISE

Ubuntu OS 22.04.1 LTS

Intel i5-10th Gen Processor 8-Core CPU

16-GB RAM

256GB SSD Storage

AZURE CLOUD

Ubuntu 20.04.4 LTS

Intel Xeon Platinum Processor 8-Core CPU

16-GB RAM

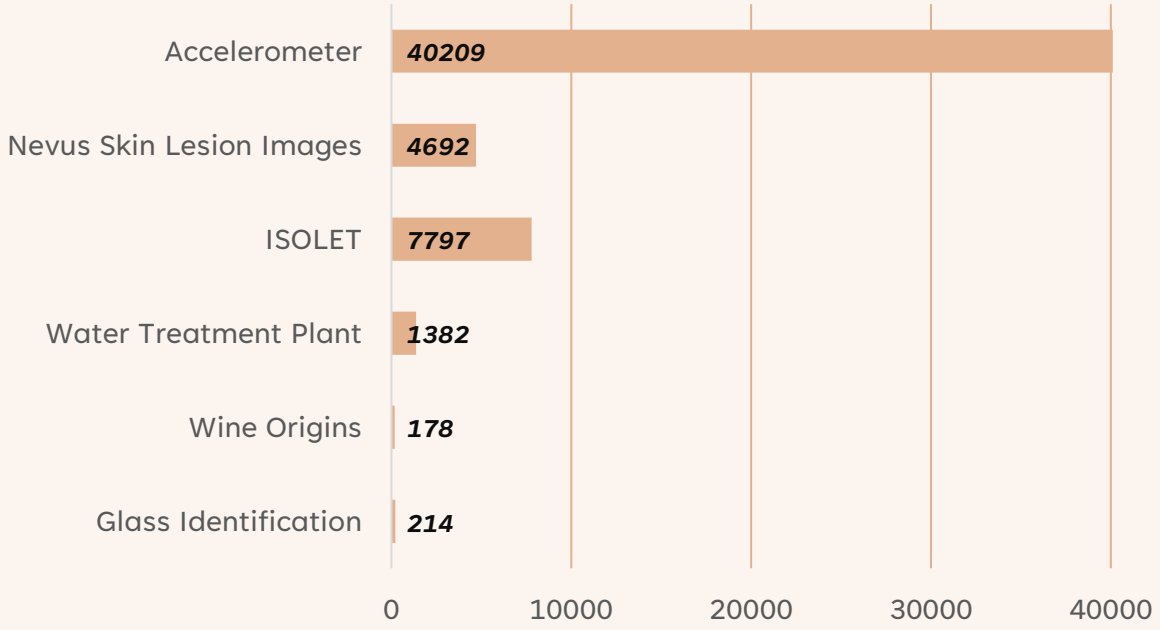
64GB Storage

WORKING WITH DIFFERENT DATA NEEDS

Considering Large & Small Datasets from different domains

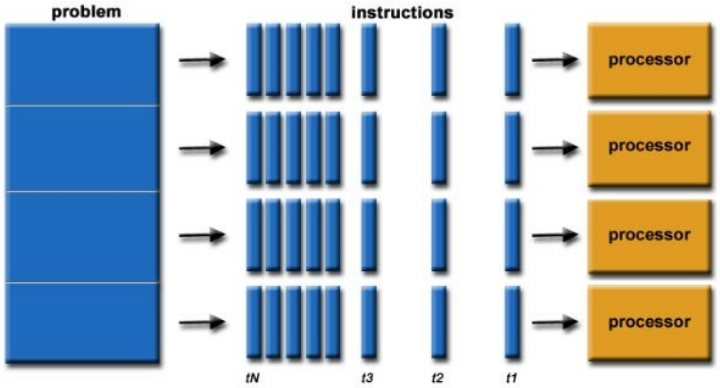
DATASET	OBSERVATIONS	ATTRIBUTES	DOMAIN
Glass Identification	214	10	Chemistry
Wine Origins	178	13	Chemistry
Water Treatment Plant	1382	19	Environment
ISOLET	7797	618	Technology
Nevus Skin Lesion Images	4692	784	Public Health
Accelerometer	40, 209	5	Technology

DATASET SIZE COMPARISON



PARALLELIZED EXECUTION

Maximizing Core Utilization



PARALLEL EXECUTION APPROACHES

Python built-in multiprocessing module to enable process-parallelism

Scikit-Learn K-Means implementation provides OpenMP-based mechanism for shared-memory multiprocessing

EVALUATION METRICS

CPU Cores v/s Execution Time

CPU Cores v/s CPU Utilization

Comparing performance for different combinations of:

- Number of Clusters (k) – 2, 4, 6, 8
- Datasets
- Parallelization Approach

EXISTING HARDWARE OVER CLOUD SPENDINGS

Outcome of Comparison between Azure & On-Premise Systems

UTILIZING THE EXISTING INFRASTRUCTURE

The result of 4536 trials on each environment indicated that a moderately strong existing on-premise infrastructure provides fairly good performance relative to Cloud

COST SAVINGS

Need to spend extra dollars to get better performance on cloud, a potential deal breaker for small enterprises, students & researchers with existing feasible hardware

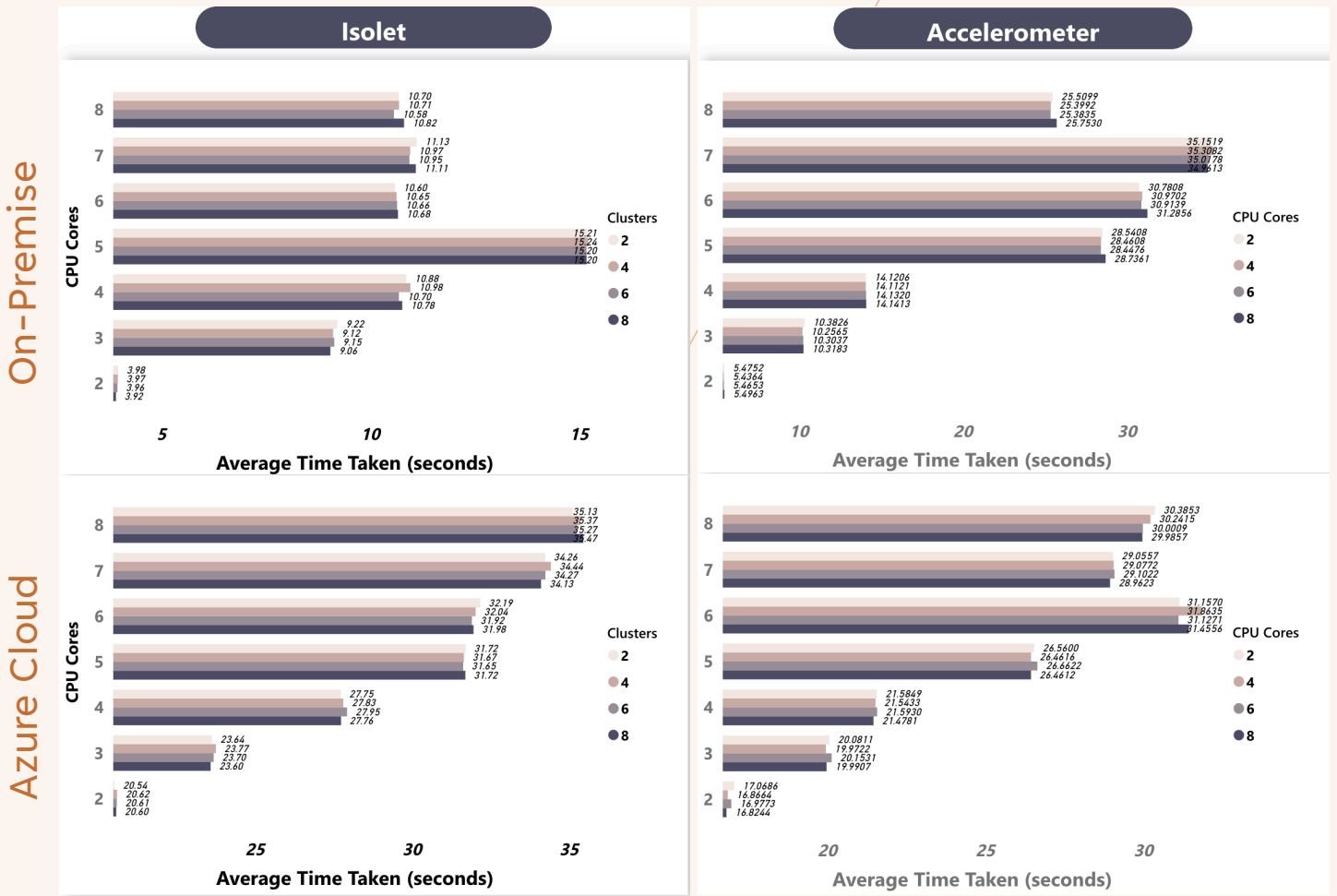
Average Execution Time

# CPU Cores	Azure	Local
2	4.92	2.40
4	4.92	2.37
6	4.92	2.51
8	4.91	2.69
Average Time in Seconds	***4.92***	***2.49***

UNEXPECTED TRENDS IN PROCESSING TIME

Abstracted Implementation Of Complexity Resulting in Uncontrolled Core Utilization

- Execution time did not decrease with increase in number of CPU cores
- Intermittent patterns of execution duration appeared in both cloud-based and on-premise environments and regardless of dataset sizes
- CPU Usages indicates under utilization of available Computation power since it doesn't increase dramatically
- No common thread to explain circumstances in which the increased number of cores resulted in prolonged processing times or decreased CPU usage
- Lack of Control over embedded implementation of K-Means and its variants obscuring the cause of unexpected trends



EXISTING METHODS RELIABILITY

Conclusions...

SCIKIT-LEARN

Although Scikit-Learn provides out-of-the- box parallelism capability that must reduce the processing times on high number of CPU cores, the research outcomes indicate otherwise

DATA SCIENCE V/S COMPUTER SCIENCE

Investigation of this irregularity requires a perspective of a computer scientist rather than a data scientist to make best use of available hardware through optimal software

REINVENTING THE WHEEL V/S ACCESSIBILITY

People working with Data need to focus on analysis & insights and thus, accessible, reliable & optimized software to work rather than worrying about the software optimization itself

INNOVATING OPTIMIZED APPROACHES

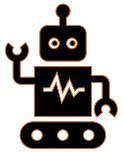
MORE EXPERIMENTATION & INVESTIGATION

MOTIVATION TO INNOVATE

This Research provides us with a motivation to innovate more streamlined clustering implementation that are optimized for varying needs & infrastructure

EXPERIMENTATION IS THE KEY

Challenging the existing approaches through experimentations in varying environments is the key to innovate better approaches, hence, a need of rigrous investigation through trials.



IMPLEMENTING MODERN TECHNIQUES

Advancements in computing over the years such as quantum computing & more, can be put to use for implementing modern solutions to clustering problems

THANK YOU!