PROPOSAL ABSTRACT

ALGORITHMS OPTIMIZATION AND DIGITAL CARBON FOOTPRINT REDUCTION BY APPLYING HIGH PERFORMANCE COMPUTING TECHNIQUES IN THE DESIGN OF COMPACT DATA STRUCTURES FOR PROBLEMS IN THE BIG DATA DOMAIN.

There is no rigorous definition of big data. Initially the idea was that the volume of information had grown so large that the quantity being examined no longer fit into main memory for their processing, so engineers needed to change the way they used for analyzing it all. Today, the specific problem is that we are overwhelmed by not having the ability to process/analyze such a magnitude of new information that we receive second by second. Therefore, new research fields such as Data Science (DS) have deepened their efforts to contribute to big data problems. In short, DS is a multidisciplinary blend of data inference, algorithms development and technology in order to extract useful information, for a specific business process, from analytically complex data. However, there are still many problems associated with data manipulation, being the greatest limitation of the sophisticated software tools for big data their approach to obtain critical information, which is based on searching directly over large sets of raw data. Since working with unstructured data without preprocessing to filter or normalize the input data sets will lead us to run expensive routines in terms of CPU time, memory, and power consumption. The effect in the scientific community is that many researchers are trying to develop smart strategies to infer information on big data. They are looking for new ways to discover critical business information implicitly stored and distributed in large-scale volumes of data, giving rise to a new research field called Big Data driven to Data Science.

In this project we will provide the theoretical foundations to extend the use of compact data structures to do data science on big data. The proposal includes to investigate how it is possible to use CPU/GPU architectures, and others High Performance Computing (HPC) techniques, in Data Compression (DC) algorithms and how that can help to accelerate final user queries on Compact Data Structures[1](CDS) designed for data science, or to solve a specific problem for big data. Our main objective is to reduce the digital carbon footprint by speeding up the execution times of algorithms for big data, through HPC optimizations on: *i*- data compression algorithms and *ii*- the construction and use of compressed/succinct structures that replace the input, thus facilitating the subsequent operations involved in the process, be it data science or to solve a specific problem in the big data domain. To do that, our research team consisted largely of students from the area, first will investigate deeply in the fields of DC and CDS design to propose new contribution based on HPC. Then with our basic results, we implement our main methodology that we summarize in Figure 1. Our scheme simplifies each problem in big data by executing two stages: a preliminary stage to process the raw data, building intermediate structures that are smaller than the input, in order to solve the problem more efficiently in a second stage by using these built structures instead of working directly on the raw data.

Our approach address real-world problems, bringing new life into old problems, and completely changing the perspective from which big data driven to data science is being done today; where current proposals have focused mainly on practical issues, but used the same old ideas of storing all the data and try to make science directly on them.

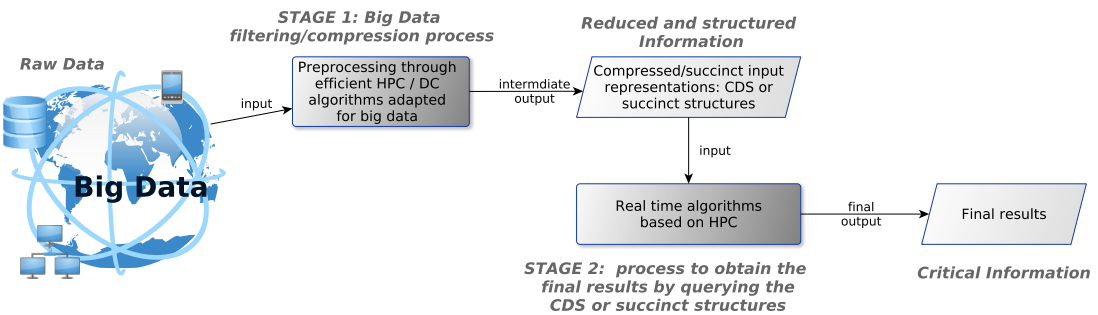

Figure 1: Illustration of the main methodology of our proposal. In a first stage, we process the huge input data, by using HPC techniques and DC algorithms, in order to build intermediate structures without redundancy or including only the useful data to solve the problem. After that, in a second stage, by using HPC algorithms designed to work on these succinct structures, we carry out a final process to obtain the desired results, which can be in data science or solving specific problems.

[1]A compact data structure is one that can replace raw input data and provides efficient operations for both extraction and computation or real-time searches on the compressed data.