Extended Abstract

Review and Research: Journal of Engineering And Technology

2021

Vol. 10 Issue 7

High-performance data analytics: Platforms, resource management and middleware

Shikharesh Majumdar

Carleton University, Canada

Abstract:

Enterprises, social networks and smart systems that leverage the Internet of Things technology often lead to large datasets. Data analytics concerns the extraction of knowledge from such raw data. The challenges underlying the processing of such data sets are captured in the 3V characteristics of Big Data: Volume, Velocity and Variety. The first refers to the large size of stored data sets, the second to data in motion streaming from social networks or sensor-based smart systems for example while the third concerns the large variety in data types and formats. High-performance computing platforms such as clusters and clouds are often deployed to address these challenges. Enabling technology that includes parallel processing frameworks and platforms, as well as algorithms for the management of resources in the cloud/cluster, is crucial for performing data analytics in a timely manner. Focusing on such enabling technology this talk will address the various challenges and potential solutions in the context of cloud-based systems for supporting Big Data analytics and smart systems. Issues to be discussed include (a) Management of resources in the context of latency-sensitive data analytics applications such as deadline driven Map Reduce jobs and mobile object tracking (video analytics) algorithms. (b) Scheduling techniques for supporting streaming data analytics. (c) Edge-computing based platforms for performing complex event processing in the context of sensor-based streaming applications such as remote patient monitoring. (d) A cloud-based middleware for the unification of geographically dispersed resources required in the management of smart systems such as sensor-based bridges and aerospace machinery. Our project is at Interface Big Data and HPC -- High-Performance Big Data computing and this paper describes collaboration between 7 collaborating Universities at Arizona State, Indiana (lead), Kansas, Rutgers, Stony Brook, Virginia Tech, and Utah. It addresses the intersection of High-performance and Big Data computing with several different application areas or communities driving the requirements for software systems and algorithms. We describe the base architecture including the HPC-ABDS; High-Performance Computing enhanced Apache Big Data Stack, and an application use case study identifying key features that determine software and algorithm requirements. We summarize middleware including Harp-DAAL collective communication layer, Twister2 Big Data toolkit and pilot jobs. Then we present the SPIDAL Scalable Parallel Interoperable Data Analytics Library and our work for it in core machine-learning, image processing and the application communities, Network science, Polar Science, Bimolecular Simulations, Pathology and Spatial systems. We describe basic algorithms and their integration in end-to-end use cases. Many scientific problems depend on the ability to analyze and compute on large amounts of data. This analysis often does not scale well; its effectiveness is hampered by the increasing volume, variety and rate of change (velocity) of big data. This project will design, develop and implement building blocks that enable a fundamental improvement in the ability to support data intensive analysis on a broad range of cyber infrastructure, including that supported by NSF for the scientific community. The project will integrate features of traditional high-performance computing, such as scientific libraries, communication and resource management middleware, with the rich set of capabilities found in the commercial Big Data ecosystem. The latter includes many important software systems such as Hadoop, available from the Apache open source community. Collaboration between university teams at Arizona, Emory, Indiana (lead), Kansas, Rutgers, Virginia Tech, and Utah provides the broad expertise needed to design and successfully execute the project. The project will engage scientists and educators with annual workshops and activities at discipline-specific meetings, both to gather requirements for and feedback on its software. It will include under-represented communities with summer experiences, and will develop curriculum modules that include demonstrations built as 'Data Analytics as a Service.'

This work is partly presented at 7th International Conference on Big Data Analytics & Data Mining | September 26-27, 2018 | Chicago, USA