Scholar's Name: Samiya Khan

Supervisor's Name: Dr. Mansaf Alam

Department: Department of Computer Science (Faculty of Natural Sciences)

Thesis Title: A Cloud Based Framework for Big Data Analytics

Abstract

Data deluge is a growing technological challenge in this era of digitization. Acquisition, storage, processing and visualization of big data for specific analytical applications have emerged as potential research areas of interest for the scientific community. Cloud computing is identified as one of the best infrastructural solutions for implementation of big data applications because of its cost effectiveness and functional efficacy.

However, synergistic use of cloud computing with the big data paradigm for development of domain-specific applications is not short of challenges. Technology and application-specific research issues exist at different phases of the big data lifecycle, which need to be alleviated for effectual applications' development. Moreover, the wide realm of applicability of big data analytics makes it a commercially potent technological field with expansive scope for interdisciplinary research.

Major research challenges targeted in this thesis include application specificity of cloud-based big data solutions and the lack of API-based solutions for domain-specific applications. This thesis identifies education and research as high-potential big datasets and application areas and proposes the concept of educational intelligence, which adapts business intelligence to develop specific applications. Moreover, it presents a detailed survey on scholarly data and potential big data applications in this field of research.

With regard to scholarly applications, this thesis focuses on the challenges posed by data preprocessing during development of deep learning applications. Thus, it proposes, implements and validates a preprocessing approach that uses Spark ML for pipelining preprocessing tasks; thereby, reducing execution time and associated costs. Furthermore, this thesis proposes a technological framework for development of educational intelligence applications and implements two applications as case studies namely, PABED, which is an analytical tool for big education data analysis and an outcome-based quality assessment framework for higher education systems.

Finally, the thesis proposes a generic framework for provisioning Big Data-as-a-Service and provides a technology selection criterion for choosing appropriate storage and processing solutions for a big data application. In unison with this, the thesis proposes a computing model-based taxonomy for big data processing solutions and clustering analysis-based classification scheme for NoSQL solutions, which is the most popular storage technology for big data systems.

Bivariate, cluster and suitability analysis of 80 NoSQL solutions have been performed. A decision tree-based predictive model for testing suitability of a NoSQL solution for an application area has been proposed and the model is available for use at www.p-nasa.com. The selection criterion can be cumulatively used for designing of custom big data stack for applications’ development based on functional and business requirements of the system.