Summer 2019

Development of HU Cloud-based Spark Applications for Streaming Data Analytics

Sangwhan Cha

Computer and Information Science
Undergraduate Project Topics and Ideas

Mina Gabriel,
CISC Experiential Learning Coordinator
Harrisburg University
326 Market St,
Harrisburg, PA 17101
(717) 265-3727
MGabriel@HarrisburgU.edu
http://harrisburgu.edu/

Title:
Development of HU Cloud-based Spark Applications for Streaming Data Analytics

Author:
Sangwhan Cha - scha@harrisburgu.edu

Difficulty:
Hard

Specialization:
Artificial Intelligence

If other, please specify:

Most Appropriate Course:
Project II

Brief Description:
Nowadays, streaming data overflows from various sources and technologies such as Internet of Things (IoT), making conventional data analytics methods unsuitable to manage the latency of data processing relative to the growing demand for high processing speed and algorithmically scalability [1]. Real-time streaming data analytics, which processes data while it is in motion, is required to allow many organizations to analyze streaming data effectively and efficiently for being more active in their strategies. To analyze real time âĂIJBigâĂİ streaming data, parallel and distributed computing over a cloud of computers has become a mainstream solution to allow scalability, resiliency to failure, and fast processing of massive data sets. Several open source data analytics frameworks have been proposed and developed for streaming data analytics successfully. Apache Spark is one such framework being developed at the University of California, Berkley and gains lots of attentions due to reducing IO by storing data in a memory and a unique data executing model. In Computer Information Sciences (CISC) at Harrisburg University (HU), we have been working on building a private Cloud Computing for future research and planning to involve industry collaboration where high volumes of real time streaming data are used to develop solutions to practical problems in industry. By developing a HU Cloud based environment for Apache Spark applications for streaming data analytics with batch processing on Hadoop Distributed File System (HDFS), we can prepare future big data era that can turn big data into beneficial actions for industry needs. This research aims to develop Spark applications supporting an entire streaming data analytics workflow, which consists of data ingestion, data analytics, data visualization and data storing. In particular, we will focus on a real time stock

recommender system based on state-of-the-art Machine Learning (ML)/Deep Learning (DL) frameworks such as mllib, TensorFlow, Apache mxnet and pytorch. The plan is to gather real time stock market data from Google/Yahoo finance data streams to build a model to predict a future stock market trend. The proposed Spark applications on the HU cloud-based architecture will give emphasis to finding time-series forcating module for a specific period, typically based on selected attributes. In addition, we will test scale-out architecture, efficient parallel processing and fault tolerance of Spark applications on the HU Cloud based HDFS. We believe that this research will bring the CISC program at HU significant competitive advantages globally.

Number of students needed:
3

Outcomes and Deliverable:
Cloud-based Spark Applications for Streaming Data Analytics

Skills Required:
Hadoop, Spark, HBase, Python/Scala/Java, Basic of ML

Available Resources:
HU Cloud system

Program Goal:
CISC 1.1: Mathematical Analysis, CISC 1.3: Develop Solution, CISC 1.4: Deploy Solution CISC 2.1: Hardware Platform, CISC 2.2: Software Platform, CISC 2.5 Analysis of AlgorithmsCISC 3.1: Explore New Methodologies, CISC 3.2: Explore New Design CISC 4.1: Written Communication, CISC 4.2: Oral Communications CICS 5.1: Collaborative Work Practices

Student Learning Outcomes:
1a: The student should be able to analyze a problem in a manner that facilitates the design of its solution., 1b: The student should be able to apply relevant principles of computing during their analysis of a problem., 1c: The student should be able to apply relevant principles of related, non-computing disciplines during their analysis of a problem., 2a: Student is able to create a formal software design based on a given set of requirements., 2b.:Student is able to develop a software solution from a formal design specification., 2c: Student is able to evaluate a software solution to determine its compliance with the specification., 3a: Student will be able write in a standardized format in order to organize their thoughts and deconstruct their ideas at a level appropriate for the desired audience., 3b: Student will be able to verbally communicate effectively with an advisor, group of collegues or an audience to express a thought or idea at a level appropriate for the desired audience., 4a: demonstrate understanding of legal and ethical principles., 4b: demonstrate working knowledge of a code of ethics., 4c: exhibit informed judgement involving legal and ethical decisions., 5a: Ability to organize tasks, contribute a fair workload, and see tasks to completion., 5b: Ability to collaborate as an effective team member., 6a: Student will be able to produce computer-based solutions by applying applicable computer science theory and software development fundamentals