Fall 2019 Term

Project 1

Title: Erasure Coding for HDFS

Course: CS 1950

Abstract: Hadoop Distributed File System (HDFS) traditionally uses 3x replication when storing data in order to provide fault tolerance. Erasure coding is an alternative to replication that provides flexibility in selecting the degree of redundancy depending on the fault tolerance requirement. In this project, the student will explore the viability of erasure coding by adding such functionality in Libhdfs3, a library that allows clients to connect to a Hadoop cluster and perform file system operations on HDFS, and subsequently comparing the implemented functionality with traditional replication in terms of performance as well as resilience support.

Required Skills: C++, familiarity with multi-threaded programming

Team Size: 2 students

PittCS Supervisor: Panos K. Chrysanthis

Vertica Supervisor: Deepak Majeti

Location: ADMT Lab and Vertica, Pittsburgh

Project 2

Title: Evaluating different HDFS C++ clients

Course: CS 1980 / CS 1950

Abstract: Libhdfs++ and LibHdfs3 are C++ libraries that allow clients to connect to a Hadoop cluster and perform file system operations on HDFS. Furthermore, WebHdfs is a Rest Api that performs the same operations via http. The goal of this project is to compare these different mechanisms in terms of the functionality they support, their performance, and also their ability to handle large concurrent requests. Students will study the code to understand the differences in the libraries' implementations, and also design experiments to compare the performance and scalability of these mechanisms.

Required Skills: C++, familiarity with multi-threaded programming

Team Size: 3 students

PittCS Supervisor: Panos K. Chrysanthis and Constantinos Costa

Vertica Supervisor: Deepak Majeti

Location: ADMT Lab and Vertica, Pittsburgh

Project 3

Title: Optimizing Apache Parquet Reads for S3

Course: CS 1950

Abstract: Apache Parquet (https://parquet.apache.org/) is a popular columnar file format used for storing and querying big data. Parquet file format is supported by all major analytical tools such as Hive, Spark, Presto, Vertica, etc. Object store file-systems (such as Amazon's S3) have recently become popular to store data. The cost of reading and writing data to S3 depends on the number of API calls made. The scope of this project is to optimize SQL analytical queries by tuning execution performance and cost of reading/writing data to S3.

Required Skills: C++, Knowledge of Object Stores such as Amazon S3

Team Size: 2 students

PittCS Supervisor: Panos K. Chrysanthis

Vertica Supervisor: Deepak Majeti

Location: ADMT Lab and Vertica, Pittsburgh

Project 4

Title: Apache Parquet Column Indexes

Course: CS 1950

Abstract: Apache Parquet (https://parquet.apache.org/) is a popular columnar file format used for storing and querying big data. Parquet file format is supported by all major analytical tools such as Hive, Spark, Presto, Vertica, etc. The Parquet specification recently introduced Column Indexes to allow fast data lookup. This project involves implementing and investigating the Column Indexes inside the Apache Parquet C++ library. The investigation will include optimizing Column Indexes for memory and performance when writing and reading parquet files.

Required Skills: C++

Team Size: 2 students

PittCS Supervisor: Panos K. Chrysanthis

Vertica Supervisor: Anatoli Shein

Location: ADMT Lab and Vertica, Pittsburgh

Project 5

Title: Backup and Restore on HDFS

Course: CS 1950

Abstract: Backup and restore are essential database components needed to recover a database from physical failures such as storage or operating system crashes. Hadoop Distributed File System (HDFS) is a popular file-system used by many large enterprises to store cold data. HDFS supports data fault-tolerance and is well-suited to be a backup location.

The scope of this project is to implement backup and restore operations in Python that help synchronize data and catalog objects of an entire Vertica database to HDFS. The project will include the performance evaluation and optimization of these backup/restore operations.

Required Skills: Python, C++, familiarity with multi-threaded programming

Team Size: 2 students

PittCS Supervisor: Panos K. Chrysanthis

Vertica Supervisor: Jie Guo

Location: ADMT Lab and Vertica, Pittsburgh

Project 6

Title: Context-Aware Path Recommendation

Course: CS 1950 or CS1980

Abstract: During extreme weather conditions and natural disasters caused by meteorological phenomena, it is imperative to enable navigation that minimizes the outdoor section of recommended paths. CAPRIO (http://db.cs.pitt.edu/caprio) is a context-aware path recommendation system whose objectives are two-fold: (i) minimizing outdoor exposure; and (ii) minimizing the distance of the recommended path. In this project, the student(s) will extend the system to integrate more information about the potential path (e.g., the accessibility of the building) using the graph integrator module to enrich the quality of the recommended path. The student(s) will also explore the CAPRIO's data management subsystem to improve the overall robustness and performance of the system.

Required Skills: JAVA, familiarity with RESTFul API, HTML and Javascript

Team Size: 2 students

Supervisors: Panos K. Chrysanthis, Constantinos Costa

Location: ADMT Lab