Project Ideas-Suggestions-Directions

Cloud Computing

An assembled list of initial project ideas can be found below. **I am still editing and expanding this list, so anyone of you can freely suggest directions for the rest of the students even if you are going to be working on something else.** If you have a cool idea please let me know and I will include it in this list! You are not confined to follow any of the directions proposed below. You may select any project related to cloud computing that interests you, after discussing it with me and getting approval of course.

Parallel Computing with Cloud -- Potential Platforms:

Cloud computing engines such as Spark and Hadoop and many other related engines and libraries facilitate parallel computing on Cloud.

For your projects you can use:

1) the Docker Containers distributed to students for Spark and Hadoop.

2) You may open accounts to our own Amarel Lab, log in remotely, and be able to use Spark and many other tools for your project.
https://oarc.rutgers.edu/resources/amarel/#access

3) You can use Elastic Map Reduce (https://aws.amazon.com/elasticmapreduce/) from amazon to start experimenting with Hadoop and Spark. Both of these tools are installed on EMR and by reading through amazon tutorials you will learn how to launch a cluster on EMR and start your project with Hadoop or Spark. It is preferable that you first install these engines on your local machine, you can follow the below links to do so:
For **hadoop**, http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
For **spark**, http://spark.apache.org/docs/latest/building-spark.html
Be careful, to build spark, you should have Maven 3.3.3 or newer and Java 7+."

You may also visit the following links:
https://aws.amazon.com/articles/Elastic-MapReduce/4926593393724923
http://ampcamp.berkeley.edu/big-data-mini-course/
http://people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf

https://www.engpaper.com/2019-papers.htm

FINAL PROJECT PROPOSAL SUGGESTION 1

To decide what kind of problem, application, or numerical method you like to scale on EMR using Hadoop or Spark please read papers from the conferences listed in:

https://www.engpaper.com/cse/index.html DATA BASE OR RECENT AND RELATED CONFS.

High Performance Computing Conference: https://hipc.org/

IEEE Int'l Conference on Cloud Networking: https://cloudnet2022.ieee-cloudnet.org/

IEEE Cloud Conference: https://conferences.computer.org/cloud/2023/

SuperComputing Conference: SC: http://sc22.supercomputing.org

Int'l Parallel and Distributed Symposium IPDPS: https://www.ipdps.org/ipdps2023/2023-advance-program.html

Symposium on Operating System Principles: SOSP: https://sosp2021.mpi-sws.org/program.html

USENIX Symposium on Operating Systems: OSDI: https://www.usenix.org/conference/osdi23/technical-sessions

Int'l Conference on Very Large Databases VLDB: https://vldb.org/2023/?program-schedule

Cluster, Cloud, and Internet Computing CCGIRD: https://ccgrid2023.iisc.ac.in/

You may also work with implementing one of the papers titled "machine learning" in the paper presentation folder I have provided for the course.

FINAL PROJECT PROPOSAL SUGGESTIONS 2

Traditional Parallel Computing vs Parallel Computing with Cloud: Take a scientific computing problem and implement it using Spark and/or Hadoop (resources for Spark and Hadoop on AWS are documented above). Then implement the same problem on a cluster of EC2 nodes from AWS using MPI (follow the resources in class slides to learn MPI and additional resources). Compare the two implementations and clearly identify why you see different performance results. You can then research on why one platform is better and how your algorithms on either of the platforms could be modified to run faster and more efficiently.

GPU Virtualization on Cloud and Mobile Cloud: Read the following three points from "Eyal de Lara" to get some ideas on what you can do in this area:

GPU accelerate memory encryption for smart phones: Smartphones are ever more popular. Their growing presence in e-commence, mobile payment, healthcare, etc. requires sensitive data to be stored on the phone. Such data, in the event of phone lost or theft, could be maliciously exploited if not protected. Today popular smartphone OS, such as Android, however, lacks a robust mechanism to prevent such data theft. This project explores the development of a secure smartphone that automatically encrypts itself when the phone goes to sleep and decrypts itself when the phone resumes. The project will leverage the phone's GPU to accelerate the computationally intensive data encryption and decryption.

GPU Acceleration of Cloud Management Tasks. Modern CPUs, such as AMD's Fusion architecture, integrate a GPU into the main processor die. We are researching the use of the GPU to optimize system level tasks in cloud computing infrastructure. Some possible uses of the GPU include zeroing pages, digest generation to reduce the overhead of memory deduplication (a technique that reduces memory pressure by identifying common pages in co-located VMs that can be shared), or to support content addressable storage (CAS), memory compression, and parallelism the page table remapping step, a key stage in the process of suspending and resuming a VM in which the physical frames are reassigned to new machine frames. A course project in this space could implement and evaluate the benefits of one such algorithms, e.g., memory compression/decompression.

Next Generation GPU Virtualization The project explores the use of GPUs in virtualized cloud environments. Modern GPUs, such as nVidia's Grid K1, provide hardware support for virtualization which approximates native performance by running the GPU driver inside the virtual machine (VM). Unfortunately, this approach currently limits prevents VMs from being migrated

and does not support the dynamic reassignment of GPU resources. This project explores a new architecture which achieves near-native performance while preserving the flexibility to migrate VMs and reassign GPU resources.

Mobile Cloud. There is enormous potential for context-rich sensing on mobile devices in support of quality-of-life enhancing applications, which for example exploit video processing to perform activity recognition, or process speech to perform automatic translation. Whereas, future mobile devices will have the sensing capabilities required for these tasks, form factor concerns will limit their computing and storage capabilities. Cloud computing, where a third party, such as Amazon, assumes the costs of provisioning and operating the data center, and offers computation and storage on demand as a metered commodity, represents an attractive environment for deploying future high performance pervasive applications. One possible projects in this space would use the Hawaii toolkit in combination with Microsoft Azure to create a next generation context-aware mobile application.

FINAL PROJECT PROPOSAL SUGGESTION 3

Contributing to Spark, Hadoop, and other Cloud Engines: Spark and Hadoop are OpenSource. There are missing parts in these engines and researchers around the world are actively contributing to these tools. For example read about MLIB: http://spark.apache.org/mllib/ or GraphX: http://spark.apache.org/graphx/ Both of these libraries are missing machine learning algorithms and graph analysis methods. Document your contribution for your project proposal.

Disease Study with Cloud: Cloud computing is built to analyze massive amounts of data. Many researchers are using cloud to detect faulty genes in the human genome. Some are mapping brain activity with cloud. Here are some good links to papers and tools:

Papers on genome research on cloud: http://www.cc.gatech.edu/~saluru/publications.html
Brain Activity:
http://people.csail.mit.edu/matei/courses/2015/6.S897/readings/freeman-2014.pdf
Neuroscience and cloud:
http://people.csail.mit.edu/matei/courses/2015/6.S897/readings/freeman-open-source.pdf
Thunder: http://thunder-project.org/

FINAL PROJECT PROPOSAL SUGGESTION 4

PROJECT IDEAS FROM PREVIOUS STUDENTS:

Fall 2023

1) **Crime Analysis with Pandas over Spark**

The aim of this project is to take advantage of this new pandas-over-spark API by providing a cohesive view of all crime statistics in a given area through the use of larger datasets that will be preprocessed and feature engineered in a distributed fashion. The motivation for this project stems from the need to know crime behavior in states, as graduating seniors are moving to accommodate for their work opportunities. As such, it is important to know if the area is safe or safe enough when considering areas for relocation.

For the project, we used New York City's crime statistics, publicly available through the NYPD. This dataset includes date, time, location, type of crime, if an arrest was made and the crime was armed. Using this dataset, we compile different statistics to display to the user. Some of these statistics include the total number of arrests made for a given offense type, which hour of the day certain crimes occur, and a showcase of most popular crime areas onto a NY map.

2) **Real-Time Data Streaming and Stock Data Analysis using AWS**

Analysis of Stock Data: We have tried to combine advanced cloud technologies with financial analysis, providing a comprehensive and insightful look into the stock performance of some of the world's leading tech giants, especially during periods of economic uncertainty and recession. This section thoroughly examines the historical stock data of 'FANGMANT' companies. By leveraging a combination of AWS services such as S3, Glue, Athena, and QuickSight, we aim to dissect and understand these companies' market behaviors and financial patterns. The focus is on generating statistical insights and interpreting the economic implications during various market phases, including periods of economic uncertainty and recessions.

2. **Real-Time Streaming**: The second section pivots to the technical implementation and challenges of real-time data streaming. Utilizing AWS's potent tools like Kafka, S3, and EC2, we aim to simulate a real-time stock market environment. This involves streaming live data, managing it efficiently in a cloud environment, and analyzing it in real-time. This part of the project underscores AWS's technological prowess and adaptability in handling dynamic, high-volume data streams typical in financial markets.

3) **K-Means Algorithm: Analysis in Parallel: Local vs. Cloud**

Today, we live in a world full of data, with everything being online; data such as website analytics, surveys, customer trends, etc, is everywhere. Modern businesses have become more reliant on this data, and processing it efficiently, making conclusions, and executing business ventures on those conclusions have become ever more critical. However, not all businesses can house vast amounts of data or perform useful

transformations on it. As such, cloud-computing-as-a-service has increased, allowing businesses to offload the low-level, behind-the-scenes operations that make data analysis happen. However, we wanted to know if this is the optimal solution and what it would have been like if a business had decided to run the computations themselves, being limited in pure machine processing power. As such, this project is about Traditional Parallel Computing vs Parallel Computing with Cloud. In this scenario, we used Google's BigQuery to emulate a business outsourcing their data work to a cloud provider vs running such tasks on a local computer. We will also look into single vs multi-threaded performance for a more holistic view.

Background: The algorithm we settled on using for our test is the K-means algorithm. This algorithm was created in 1957 by Stuart Lloyd as a pulse-code modulation technique, a method used to represent sampled analog signals digitally. He found that he was required to organize the signal space in clusters, so he proposed this algorithm.

4) **Valuation of the Forecasting Stock Prices of Publicly Traded Companies Using Spark Framework**

In the volatile landscape of financial markets, precision in stock price forecasting and intrinsic value assessment stands as a paramount necessity.

This project endeavors to establish a resilient framework for the meticulous valuation and forecasting of publicly traded entities, harnessing the prowess of the Apache Spark framework for unparalleled efficiency in data processing and analysis.

- **Goal 1:** Develop a comprehensive and sophisticated framework for the valuation and forecasting of publicly traded companies, integrating advanced analytical methodologies.
- **Goal 2:** Apply technical and quantitative analyses to dissect and interpret intricate patterns within stock market data.
- **Goal 3:** Implement a robust data preprocessing and feature engineering pipeline to curate historical and real-time financial data for optimal model training and analysis.

Fall 2022

1) **Python versus PySpark Showdown**

What it would take to run a Spark application with Stock Python? Why to use Spark? How is performance affected by a number of workers, data size etc.? The students ran experiment parallel code using both Spark and Python and compared the results. Also, in order to show the merits of Distributed Computing on the Cloud, students used Amazon S3 Bucket and AWS Glue Studio. The results demonstrate the superiority of pySpark over Python for an increasing amount of workers and increasing input to process.

2) **PageRank with Spark and MPI (also described below)**
3) **Twitter Sentiment Analysis (also described below)**
4) **Tweet Sentiment Analysis on Movies**
5) **Low Latency Distributed File System**
   The students designed a distributed file system with can function as an attached storage, with low latency as the main focus. They built a structure similar to Google File System that was heavily simplified, yet it performed the basic operations that GFS does.
6) **Real Time Anomaly Detection using Spark and Kafka**.
   The primary goal was to build an anomaly detection model to detect outliers in time series datasets. Applied methods like isolation forest to do this. Isolation forest can work with both unsupervised and supervised type problems. The dataset used was the SKAB dataset which is one of the benchmark dataset for unsupervised anomaly detection.
7) **Network Traffic Analyzer**
   Increase in network attacks lead to the increase of network security tools needed to protect against them. Network traffic analyzer tools can detect attacks before it' not too late and prevent attacks from happening. Network analyzers are now a must in every environment. Network traffic analysis and classification conducted using Spark and MLlib. Compared different Machine learning models to determine efficiency and accuracy and evaluate the performance. Determined the optimal model to achieve target results. The network analyzer uses Spark and the Machine learning library (MLlib) to train and test network data. Two data sets were used, the CICDDOS2019 and CICIDS2017 data sets from the Canadian Institute of cybersecurity. The analyzer runs on an AWS EC2 instance with 2 vCPUs and 8 GBs of memory, in a docker container that runs on the instance. Spark session environment is local using two threads (one for each CPU). Logistic regression offered the best results in terms of time and accuracy. If time isn't critical, for example analyzing static older data, the Decision tree offers the most accurate results.

Fall 2021

1)) Asyncronous Deep Reinforcement Learning - How parallelization can improve nonlinear control

2)) Movie Recommendation System

3)) Emotions classification on Cloud using SparkTorch library: "*This project was inspired by disease studying applications on cloud and is using SparkTorch library to perform classification tasks on raw data of human emotions. Deep learning application with cloud has the following advantages: 1. It enables large scale*

computing, making application highly scalable. 2. The cluster settings make hardware configurations highly reliable. 3. Branching off from the first advantage, data sharing and computation resource sharing is a huge advantage for deep learning application, as it requires extensive computation and storage demands. Emotion Classification is a prominent application in the field of human brain studies. It's non-intrusive which lessen the limitation for comprehensive studies. Detecting one's emotion offers extra tangible information to be derived that does not depend on verbal communication."

4)) Disjoint set Union Implementatation using Hadoop and Spark: "*For this project, students implemented a disjoint set union data structure which keeps track of the connected components of the graph. Theyimplemented the algorithm using Hadoop as well as PySpark. Student came up with a hypothesis intertwining several factors, and then used that to compare and contrast the results obtained. Motivation: The motivation for this paper was to increase the DSU performance by using the MapReduce paradigm. With Big Data being an ever growing field, DSU would benefit greatly with increased performance and would open a lot of doors for distributed implementation on the cloud.*".

5)) Kubernetes internals and demo: "*Containers have become an increasingly popular way to reduce resources and can make initial setup easier for users. Technologies such as Docker and Kuberentes have been on the forefront of implementing containers to developers projects. Kubernetes can make deployments of multiple containers a lot more cost effective in comparison to multiple virtual machines in lighter weight applications that don't need a whole lot of computing power. Kubernetes is an application orchestrator. It orchestrates cloud native microservices and other applications. This means that it can run parallel and distributed tasks on the cloud with multiple clusters. The technology was developed by Google in the late 2000's and early 2010's and has quickly been adopted in the industry with companies such as Amazon, Google, IBM, and Cisco. By being built on Docker, you get some of the same functionality but with more features and since Kubernetes is open source more features can be added. It also performs and is easier to deploy and manage compared to a Docker Swarm. Due to its increasing adoption and popularity in the industry, we wanted to provide a quick introduction to our colleagues*".

6)) Spark Docker: "*As an open source Framework, being able to use Spark clusters in such a flexible way provides a more efficient foundation as a Big Data Tool. The Docker Image Sports a higher number of features and capabilities, than other systems such as MapReduce. It can also apply greater bouts of speed. Being able to apply different APIs present possibilities that have enabled spark to move into a more mainstream technology, becoming more heavily used with the increased proliferation of Data among users. Providing a vehicle through CPU intensive tasks can be distributed and provides an easier way to build and ship different applications both large and small, representing a change in how we more effectively network and grow. In summation, the layer abstraction that the docker limit provides, can show the compatibility between different machines allowing big data to work on numerous platforms with the same general image from one engine*."

Fall 2020

Spark Streaming, Log files analysis to identify who transmits and what gets transmitted, who receives, get statics on popular music radios.

Presidential Debate Sentiment Analysis

Object classification task with Cloud Computing via pySpark

Streaming and fine tuning Spotify's recommender

Analysis of most common words by author on Project Gutenberg -- towards creating a Hadoop words library with extra features.

AES Encryption/Decryption using data parallel programming, e.g., Hadoop/Spark

PageRank Representation via Hadoop and MPI – thorough analysis and comparisons

Work with Kubernetes, presentation of the framework and implementation of a simple users' parallel computation there.

Movie Recommendation Systems.

Fall 2018, Fall 2019

ADL recognition using Spark

Distributed Machine Learning in AWS

Analysis of Iterative Algorithms in Spark

Implementing Recommender Systems on AWS multi-instances

Analysis of Travelling Salesman Problem using various Parallel Techniques

Raft Consensus Algorithm and Implementation

Movie Recommendation Service in Spark

Twitter Sentiment Analysis using PySpark and Naïve Bayes Classification

Implement and familiarize with the system implementation and evaluation proposed by Zhang et al. in the presented "Virtual Reality" paper.