15-319: COURSE PROJECT – IMAGE PROCESSING IN MAPREDUCE

Project Start Date: Jan 16th, 2012
Project Due Date: April 25th, 2012

LEARNING OBJECTIVES

In the previous project phase (I-A), you learnt the basics of MapReduce and Amazon EC2. This phase of the project will require you to design and implement a MapReduce application from scratch. You will get experience deploying a real-world data-intensive application. You will also work on characterizing the MapReduce application to learn its features and understand various techniques that can be used to optimize the runtime of the application.

BACKGROUND

Image processing is an important domain with significant applications in domains as diverse as medicine, defense, security etc. With the advent of multiple enabling technologies such as the Internet and ever increasing digitization of images, researchers are grappling with ever increasing amounts of image data that need to be processed. In this project you will implement an image processing framework in Hadoop. You should choose an image processing primitive (as an example, Figure 1 illustrates the Sobel edge detection algorithm) and implement it using the Hadoop framework.

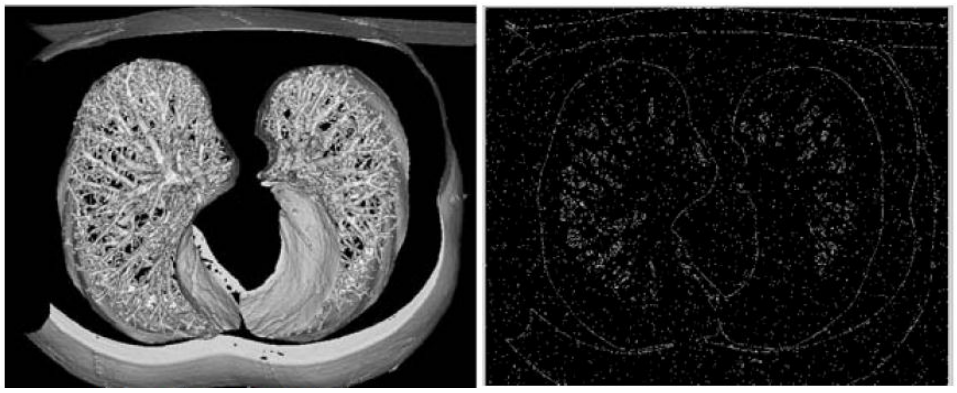

Figure 1 - Example of an Image Processing application - the Sobel edge detection algorithm applied to the source MRI image (left) yeilds an image with the edges highlighted (right)

Once you have designed and implemented the Hadoop application, you will be required to characterize it. Application characterization involves the capturing the processing and I/O characteristics of your application in order to analyze the performance of your application. Characterization is important as you can:

1. Provide insights into the Hadoop framework and discover possible bottlenecks/ intricacies, which can be used to trigger framework improvements.
2. Provide a quantitative foundation for MapReduce researchers and developers seeking validation for their hypotheses against real-world workloads.
3. Help researchers and organizations adopt sound experimental methodology by highlighting characteristics of your particular application type and fine-tuning them.

PROJECT GUIDELINES

In this project you will be working closely with a mentor that has been assigned to you. You should work closely with your mentor and schedule weekly one-on-one meetings with them.

1) Select a target image processing primitive and find a large dataset.
2) **Phase 1B**: Plan your project, state your purpose, objectives, constraints and assumptions. Present a plan with weekly deliverables in the class.
3) **Phase II:** Build a complete image processing pipeline using Hadoop MapReduce.
    a. The input to your application should be a directory containing the source images.
    b. The output from your application should be a directory containing the processed images.
4) **Phase III**: Characterize your application with respect to:
    a. Data Patterns within the MapReduce phases
    b. Concurrency of the Map and Reduce phases
    c. Phase Timelines
    d. Dataset Types
    e. Network Traffic Patterns
    f. System Resource Utilization
    g. Communication to Computation Ratio
5) Find an optimized configuration for your Hadoop application in order to minimize its runtime.

Deliverable: Write a research paper on your problem, approach, implementation, experiments, results and observations in a typical research paper format. The paper should summarize your entire project succinctly.