CSE 599c — Big Data Management Systems (Balazinska)
CSE599c Project Milestone Report

Tony Cannistra
tonycan@uw.edu

The purpose of this document is to serve as a mid-quarter update to my course project entitled "To Spark or Not To Spark: An analysis of advantage at scale." The project intends to develop rough data size thresholds for domain scientists to use when evaluating whether or not to deploy a parallel RDBMS for analysis.

1 Spark Deployment

My initial foray into this project was to evaluate the feasbility of using the Databricks Community Cloud deployment of Spark for the purposes of this experiment. After some experimentation and thought I determined that to truly perform this experiment to the extent that I was interested in, I would need to expand the cluster size beyond what Databricks offers for their cloud tier (1 master node, 6GB of RAM).

To this end I've requested AWS support from the IGERT funding to the order of $50-$100 to deploy Spark on AWS with a 4-node cluster.

2 Data Acquisition

I performed initial proof-of-concept experiments with a geospatial gridded world climate dataset (363MB, 1.2m rows) with the intent to implement a data processing pipeline for species distribution modeling, and the ease of the approach with that size data indicated that this particular benchmark would serve as an interesting test case. To flush this out more, I'm working with one of the members of my lab to acquire the necessary data scale this pipeline to 100GB of data.

3 Query Development

In my current work I'm developing a pipeline to convert climate data of arbitrary spatial and temporal resolution into bioclimatic variables useful in species distribution modeling. This approach involves aggregation, UDFs, and joins on very large tables. I think these queries will serve as a good way to compare Spark and Python DataFrames, and I already have much of the process and several different data files available.

Next steps are to prepare the data in S3, modify my current code to run in Python (from R), launch a Spark cluster, perform the benchmarking, and write the report.