Exercise 4

Time tip: Roughly 45sec to 1min per 1pt
Exercise

Q1) [6 x 3pts] Suppose you are given this task graph. Each task is amenable to perfect linear speedup on workers with BSP data-parallelism, albeit with 1 unit master overhead each before and after the actual worker execution regardless of the cluster size. Assume a homogeneous cluster.

Given the following values of T2's length and cluster size $y$ (master+workers), what is the lowest possible completion time when using any of full task-parallelism, full BSP data-parallelism, or some hybrid of both as discussed in class?

A. $x = 6; y = 2$
B. $x = 6; y = 3$
C. $x = 6; y = 5$
D. $x = 30; y = 2$
E. $x = 30; y = 3$
F. $x = 30; y = 5$
Q2) [3 x 2pts] Suppose you are given a large dataset file for ML training that is of size 120 GB. What is the lowest possible I/O cost (in GB) of each of the following feature engineering operations? Ignore final output write costs and any potential gains due to caching.
A. Quadratic (order 2) feature interactions
B. Binning a numeric feature into 10 given intervals
C. Whitening a numeric feature
D. One-hot encoding of a categorical feature (assume feature’s domain has only 5000 unique values)
Q3) [3pts] Suppose you are performing model selection for a RandomForest model. For hyper-parameter tuning, you try 3 values of number of trees and 4 values of maximum tree height. To aid your interpretability, you also explore 5 different manually created subsets of features apart from the full feature set. What is the total number of models built in this model selection workload?
Q4) [4 x 6pts] Write pseudocode (or just describe precisely) using MapReduce/Spark operations to perform the following data science operations at scale:
A. Quadratic (order 2) feature interactions
B. Binning a numeric feature with given bins
C. Whitening a numeric feature
D. One-hot encoding of a categorical feature (assume feature’s domain has only 5000 unique values and is given)
**Q5)** Suppose you are given a large dataset with 50 numeric and 9 categorical features (domain size of 50 each). The HDFS file size is 3 TB.

A. **[6pts]** Write pseudocode (or just describe precisely) using MapReduce/Spark operations to compute the correlation matrix of this dataset.

B. **[4pts]** What is the rough total disk I/O cost of the above (in TB)? Include both reads and writes of intermediate data and output.

C. **[4pts]** Briefly explain how you would scale this computation on an on-premise cluster.

D. **[4pts]** Briefly explain how you would scale this computation on AWS.