DSC 102
Systems for Scalable Analytics

Arun Kumar

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. 42; full task-par with all tasks
B. 29; T1 & T2 as task-par; T3 as full data-par on all 3 nodes
C. 15.5; T1 as full data-par on 4 nodes & T2 on 5th node as task-par with T1; T3 as full data-par on all 5 nodes
D. 60; full task-par with all tasks
E. 42; T1, T2, T3 one after another, each as full data-par on all 3 nodes
F. 24; same approach as E

Correction: 21.5; T1 on 1 node & T2 data-par on 4 nodes and task-par with T1; then T3 full data-par on 5 nodes
Exercise

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)

A. 120 GB (just one filescan read)
B. Same as A
C. 240 GB (one filescan read to compute mean and stdev; another filescan read to whiten feature values)
D. Same as A
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?

72; #trees 3 x max height 4 x #feature sets (5+1)
Exercise

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)

For Input Split, assume data is sharded tuple-wise as is common in such setting

A. Map-only job! Map() takes feature vector from tuple, performs feature interactions and emits the interacted vector with the same tuple ID.
B. Also a Map-only job. Map() takes feature value from tuple, performs binning based on given bins and emits the same tuple with same tuple ID, except this feature value is now different.
Exercise

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)

For Input Split, assume data is sharded tuple-wise as is common in such setting

C. 1 MapReduce job + 1 Map-only job. First Map() takes feature values from tuple, computes sufficient stats for mean and stdev as a 3-tuple (1, value, value squared), and emits this 3-tuple as value with a single global dummy key. Reduce() iterates over all the sufficient stats 3-tuples to put together the global mean and stdev of this feature and emits that 2-tuple as output. Second Map() job takes feature value from tuple, whitens its based on (mean,stdev) 2-tuple from prior job, and emits the same tuple with same tuple ID, except this feature value is now different
Exercise

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)

For Input Split, assume data is sharded tuple-wise as is common in such setting
D. Also a Map-only job. Map() takes feature value from tuple, performs one-hot encoding based on dictionary to map category value to new feature index, obtains the 0-1 representation for that feature (potentially sparse vector), and emits the same tuple with same tuple ID, except this feature value is now replaced with the 0-1 vector.
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.

Suppose we one-hot encode all categorical features into 50-dimensional 0-1 vectors. Then total number of numerics is 50 + 9 x 50 = 500. So, correlation matrix is of size 500 x 500, which is 250,000 cells. Even with float64, it is only 2MB. So, we will use this as our aggregation state for Mappers to send to Reducer.
Exercise

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.

A. For Input Split, assume data is sharded tuple-wise as usual. One approach to compute correlation matrix uses 2 MapReduce jobs, the first to compute the per-features means and stdevs, and the second to use those to finish the correlation computations. First Map() reads tuple, converts each categorical feature to its respective one-hot encoded vector as in Q4.D earlier to stitch together full 500-dimensional numeric vector, and computes sufficient stats as in Q4.C. First Reduce() aggregates all sufficient stats to emit 2 vectors, one with the mean of each feature and one with the stdev of each feature.
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.

A. (Continued) Second Map() reads tuple, gets the 500-dimensional numeric vector again as before and then emits as sufficient stats a 500x500 matrix representing pairwise products for the aggregation needed for the numerator of the Corr matrix formula below. Second Reduce() just adds up these individual matrices and divides all cells by total example count and the respective pairs of stdevs obtained from the first MapReduce job.

\[
corr(A, B) = E[(A - \mu_A)(B - \mu_B)]/(\sigma_A\sigma_B)
\]
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.

B. 2 filescans amount to 6 TB. Intermediate data after first job has 500x2 numbers; output is 500x500 matrix. Assuming all numbers are float64, these amount to mere ~2 MB extra. So, overall I/O cost is still roughly 6 TB.

C. Shard data on Spark cluster and write Spark/MapReduce job. Note that Dask is ruled out, since 3 TB file may not fit on single-node disk, let alone DRAM!

D. Option 1: Q5.C’s approach on EMR. Option 2: single-machine Python and remote reads from S3. Latter has less parallelism and might be very slow.