DSC 102
Systems for Scalable Analytics

Arun Kumar

Review 2

Time tip: Roughly 45sec to 1min per 1pt
Q1) [4pts] Suppose you want to build a deep net for a prediction task. You decide to explore neural architectures that have up to 4 layers, with each layer having a possible 500 or 1000 neurons. You use SGD for training and tune 3 hyper-parameters: learning rate, regularization, and batch size. Suppose you try 4 values for each hyper-parameter using the grid search method. What is the total number of models trained?
Review

Q2) You are given a large \( n \times d \) matrix on HDFS that is uniformly sharded column-wise across \( k \) workers.

A. [7pts] Write a MapReduce job (pseudocode or precise prose) to obtain the row sums vector of the matrix (aka rowSums in R). It should be scalable along the number of columns.

B. [3pts] What is the minimum possible communication cost (in big O notation of \( n, d, \) and \( k \)) to get the row sums vector using MapReduce?

(Advanced; Optional) Q3) [12pts] Given a large \( n \times d \) matrix on HDFS propose a sharding and write a MapReduce job to obtain its columnar sums vector (aka colSums in R). The layout and job should both be scalable along both the number of rows and number of columns.
Q4) [10pts] You are given a large table Ratings (RID, Stars, Date, MovieID, UserID) from a recommender system stored on HDFS. You want to know the average rating (Stars) for each movie. Propose a sharding and write a MapReduce job (pseudocode or precise prose) to obtain the above. It should be scalable along the both the number of tuples and the number of movies.
Q5) Consider the same table Ratings from Q4. Suppose you are now given it is a tuple-wise sharded file on HDFS. You aim to binarize the Stars attribute as part of your ML data preparation to cast the recommendation/ratings prediction task as binary classification.

A. [7pts] First you want to see the histogram of Stars values. Write a MapReduce job (pseudocode or precise prose) to get this. It should be scalable along the number of tuples.

B. [6pts] Now based on the histogram you decide to binarize the Stars attribute with a threshold of 3.5, i.e., Stars >= 3.5 is set to +1; Stars < 3.5 is set to -1. Write a MapReduce job (pseudocode or precise prose) to do this. It should be scalable along the number of tuples.

C. [6pts] Suppose the table was ETL’d into Spark and queryable using SparkSQL. Now do both A and B above as a single SQL query each.
Q6) [4pts] Consider a company exploring a full move to AWS for their large-scale data analytics workloads and figuring out the economics of this move.

The capital cost of getting an on-premise cluster is $300mil. The operational cost of running and maintaining the on-premise cluster is $1mil per month. The quote from AWS on the monthly cost of running their workloads on the cloud is $5mil per month. Assume neither of these monthly costs change over time.

For how long will the cloud-native approach beat the cost of the on-premise approach?
Q7) [6pts] You are given a large-scale data analytics workload that exhibits sub-linear speedup on a multi-worker cluster. Its completion time with 4 workers is 80 time units. Its speedup curve from 1 to 20 workers is a straight line with a slope of 0.75. What is the workload’s completion time on 20 workers? Note that the speedup curve's origin is at (1 worker, 1x speedup), i.e., there is no intercept.
Q8) [6pts] Consider this task graph, a forest of n pairwise dependent tasks. The work times of each task are also as listed. Note that n and k are positive integer variables, and n is a multiple of e. The parameters (a, b, c, d, e) are (1, 3, 5, 2, 8) respectively.
You are given a homogeneous cluster with n/2 nodes. What is the lowest possible completion time when using pure task-parallelism? Note that BSP or hybrid approaches are not allowed.
Q9) [5pts] You are given a database with instances of two relations R(A,B) and S(B,C,D,E), wherein B is a primary key in S and foreign key in R. The set of values of B in both relations are identical. All attributes in this database are of integer datatype (4 bytes each). All tables are stored in column store format without compression on disk. The number of pages on disk of R and S are 28 million and 4 million, respectively.

What is the rough disk I/O cost (in pages) of the following query? Exclude output write costs. Assume the DRAM cache is initially empty.

SELECT MAX(C) FROM S;
Q10) [5pts] You are given a large float64 matrix $M$ of dimensions 25 million x 10 million stored in tiled layout without compression. The tiles are squares of dimensions 2000 x 2000. They stride along both dimensions by 2000 cells starting from top left. Assume a tile’s data fits exactly on one page on disk.

What is the rough minimum disk I/O cost (in pages) of the following linear algebra computation? Exclude output write costs. Assume the DRAM cache is initially empty.

Summation of $M[1 : 5,000,000][2,000,001 : 4,000,000]$

Note that $M[i:j][k:l]$ means only rows $i$ to $j$ and columns $k$ to $l$ (ends included) are read. Row/column indices start from 1.