## 

## Q&A Parking Lot

Question from Saurabh: How do you supply the class ‘com.databricks.spark.csv’ when using not using the Databricks notebook? You can use y

Question from Muthu: In Dataframe cache, would I be able to use Memory and Disk 2 (the stuff used for serializing the data for saving on storage)? If so, I understand there may be some cost to compute. In this case would the data still be columnar store?

Question from Muthu: Count on parquet --- would this do groupBy().count.head ? Why not use stats of row count from each partition?

Question from Satish:

#1: Where do I find how much data has spilled over to the disk when we call cache or persist? (This is mostly to figure what is killing my disk space)

(btw, why do we have both persist and cache when both of them internall call the same persist api?)

#2: In spark ux, if my task has LOCALITY\_LEVEL = ANY, is that a bad thing, and how do I debug why my stage doesn't run at LOCAL level.

#3: If i am just starting to move from RDD to something more modern, would you suggest moving to dataframe or datasets?

Question: What are your tips for best handling out of memory / heap space errors? Most of my errors are trying to debug OOM and optimizing not the code, but often spark configurations.

Likes: Saurabh

Suggestion: I have seen this better handled when I increase the number of tasks (i.e, making the size of each task smaller) --- Not a TA answer :)

Question from Muthu: Spark shuffle config defaults to 200. Is this tweaked depending on the total cores available on the cluster?

Question from Muthu: When a text file (csv / tsv / flat text) is read, how does spark determine the default number of partitions on the transformed dataframe? Is it based on some sort of 32 mb chunks on the file segments?

George

There was a recent databicks video on productionising streaming apps, as part of that presentation mapWithState was covered, could you please elaborate if you have some time?

Question from Jonathan: How does Spark/Databricks handle changing partition configuration when multiple notebooks are running concurrently on the same cluster?

Question from Mark: Is there a way to compute the shortest paths for all (VertexId, VertexId) pairs inside an RDD in parallel? (The graph for my application is large, but all subgraphs are small.) Or does this have to be done iteratively?

Question from Martin: In the k-means example (or any other distance-based ML algorithm supported by Spark K-NN, naive Bayes, etc.) is it possible to also output the corresponding posterior probability (or marginal likelihood ignoring class frequencies) for a point belonging to a class (by distance to mean) calculated from the underlying density function?

Question from Deepika: What do you recommend for the situations where RDD is not too big but each partition needs large RAM for initialization/loading some indices to be used for processing each element of RDD?

Question from Deepika: Is it recommended to use Yarn/mesos over standalone cluster?

Question from Zell: Is there a way to upload the jar or py script while submitting a job on a client to the master in a standalone mode? (rather than open an http address to let the master download the jar or py script from a client)

Question from Kent: What is the standard way of handling spark streaming checkpoints when your Application code changes?

Question form Satish - Could you briefly talk about the role of driver and how executors are picked up in a spark streaming application (standalone mode) and when are they released.

What are some good spark job performance analysis tools you would recommend?

Question: What would be a good starting point or reference for Spark Unit testing?