## Q&A Parking Lot

**Question from Doug Bateman**: What is this question parking lot?

**Likes this question:** Jacob Parr, Aaron Maxwell

**Answer from Doug Bateman:** With a very large class, we’ll have students post questions here. The TAs will then help answer the questions. And the questions with the most likes will also be answered by the instructor. So put your questions here using the format below.

**Question from Doug Bateman:** Here’s a sample question.

**Likes this question:** Taggart McCurdy, Andy Konwinsky, +1 from Anonymous

**Answer from Doug Bateman:** Here’s a sample answer.

**Question from Sam Penrose**: I love the annotations of transformation vs action in 01-DA … is there any chance that information could make it into the UI somehow? In my experience users sometimes are surprised by when they are and aren’t kicking off an action.

**Question from David Barnett**: Before we’re done please show us how to import the labs into our own accounts.

**Likes this question:** Suryani

**Answer from Suryani:** most of the Labs materials are available in the thumbdrive distributed in class.

**Question from Kevin Mellott**: Why does the schema print differently when using *.printSchema*, vs the output displayed for the object instance. See below for the specific example.

pagecountsEnAllDF: org.apache.spark.sql.DataFrame = [project: string, article: string, requests: int, **bytes\_served: bigint**]

root  
 |-- project: string (nullable = true)  
 |-- article: string (nullable = true)  
 |-- requests: integer (nullable = true)  
 **|-- bytes\_served: long (nullable = true)**

**Answer from Sam Penrose:** the output displayed for the object instance is Scala REPL output. printSchema() is a method designed for human readability. (You’ll notice when you call it on a DataFrame with heavy nesting)

**Question from Peter Halliday**: Is DBFS only available for DataBricks? We use Parquet on S3 a lot, and am not sure if it’s using that behind the scenes automatically? Not sure the benefits of DBFS over interacting with s3a/s3n?

**Answer from SF:** DBFS is just a wrapper around S3 to make it easier to access. And yes, it’s only available on Databricks (use S3 directly if not on DB). There are some optimizations in place to make DBFS faster to access than S3 directly.

**Question from Mahae Koh:** Why did the “// Order by the requests column, in descending order” cell fail earlier?

**Question from BradC:** Let’s suppose we have a long sequence of transformations and then a specific SQL statement, but to understand performance you only want to time some of the transformations and the SQL statement. What is the best action to insert so you only time the specific transformations you want. And what is the best action to put on the SQL statement. Something like: 1) .take(1) 2) .collect().head.getLong(0) or 3) .count() or ???

**Question from Tim T:** what is the easiest/most reliable way to tell when a spark action triggers the re-computation of earlier stages in the DAG? For instance if I see that a stage in the DAG is being executed twice when I expect it should only be executed once, how can I tell if that's because I haven't been calling .cache() correctly, or if the data has been evicted from the cache?

**Answer from SF:** You can check the Spark UI’s DAG visualization to see colors of skipped stages. If a stage is skipped it wasn’t run. So if older stages are being re-run you can tell y the color. You can also look at the Storage UI to see if data is still in Cache.

Question 2: I can see Kafka solving more of the problem you mentioned but for the problem of backend D stream cannot process data fast enough and intermediate data store (kafka) keeps increasing backlogged packets wouldn’t Kafka also blow up eventually??

**Answer from SF:** I think Kafka has ways to deal with pressure like checkpointing to local disk. Would be best to check Kafka docs for how it handles this...

**Question from Peter Halliday:** Can you describe the process of choosing the correct value for spark.sql.shuffle.partitions?

**Answer from SF:** It defaults to 200. There are handful of things to consider. This setting is basically just the # of partitions to use in the DF after a shuffle. You ideally want a partition to host 50 - 200 MB of data. You also want a task to probably run for under 15 seconds. Just increasing it to 1,000 won’t necessarily speed up your workloads if each partition ends up with a few bytes of data (network communication costs would be too high)

**Question from Janardhan**: 1. Difference between coalesce and repartition ?  
**Answer from Ariel**: The best reference is sometimes the [source code](https://github.com/apache/spark/blob/128c29035b4e7383cc3a9a6c7a9ab6136205ac6c/core/src/main/scala/org/apache/spark/rdd/RDD.scala#L376). Coalescing down to less partitions can be done narrowly, that is, without invoking a shuffle. This is because parent partitions only need to be read on one child partition, and works because we’re reducing the number of total partitions. Repartitioning is essentially coalesce plus a shuffle step, which allows for increasing as well as decreasing the number of partitions.

2. Partition logic for a dataframe when it reads from a csv file ?

**Question from Dave:**

1. Where do I get the Kafka code you showed in class? Can you please share?
2. Can multiple dataframes write to s3 file from multiple threads at the same time?

**Question from Mahae Koh**: I notice that there are encoders for primitives but not for lists or seqs. Sometimes I have to work with lists of lists. What’s the official guidance for those use cases?

**Answer from SF:** In the future, Spark will allow you to create custom encoders.

**Question from Peter Halliday**: Can you include the Labs that you showed DataFrame and DataSets with case Classes?

**Questions from Suryani:**

1. Will Tungsten be available/running by default on Spark 1.6 onwards?

**Answer from SF:** Phase one of Tungsten is already running in Spark 1.6 (I think it appeared initially in experimental API around Spark 1.4)

2. Looking forward to Lab 12 on Streaming with Kafka example sharing, please…   
3. How do we run Spark Streaming on scheduled basis, does Spark cluster compulsory?