**(Be sure to look at Week5\_Review as well for next QC)**

What is the storage level for .cache()?

* MEMORY\_ONLY

Different ways to repartition?

* .repartition
* .coalesce (to reduce partitions)

What does it mean to run an EMR Step Execution?

* When the job is finished, the cluster terminates. If ran in cluster mode, the job remains idle

What is the Spark History Server?

* Monitoring tool that displays information about completed spark applications

What does it mean to “spill to disk” when executing spark tasks?

* Since spark runs on RAM, if there is not enough memory, the data will spill to disk memory

When during a Job do we need to pay attention to the number of partitions and adjust if necessary?

* If the number of partitions grows to be to high, or if streamed data continuously creates new partitions

What is spark.driver.memory? What about spark.executor.memory?

* The amount of memory that the driver and executor have

What is a Spark Application? Job? Stage? Task?

* Application- main function
* Jobs- Work submitted to Spark. Created when an action occurs
* Stage- Jobs are divided into stages
* Task- Each stage is divided into tasks, the smallest unit of work for Spark

What is the logical plan? The physical plan?

* The logical plan is the tree that represents both the data and schema. They depict the expected output
* The physical plan decides what type of join, sequence of execution of many methods

How many partitions does a single task work on?

- One

What’s the difference between cluster mode and client mode on YARN?

- In cluster mode, the driver runs inside an application master process which is managed by YARN on the cluster and the client can go away after initiation. In client mode, the driver runs in the client process and the application master is only used for requesting resources from YARN

What is an executor? What are executors when we run Spark on YARN?

- Executors are worker nodes’ processes in charge of running individual tasks in a given Spark job.

What is a Dataframe?

How would we go about using Spark SQL to query Dataframes?