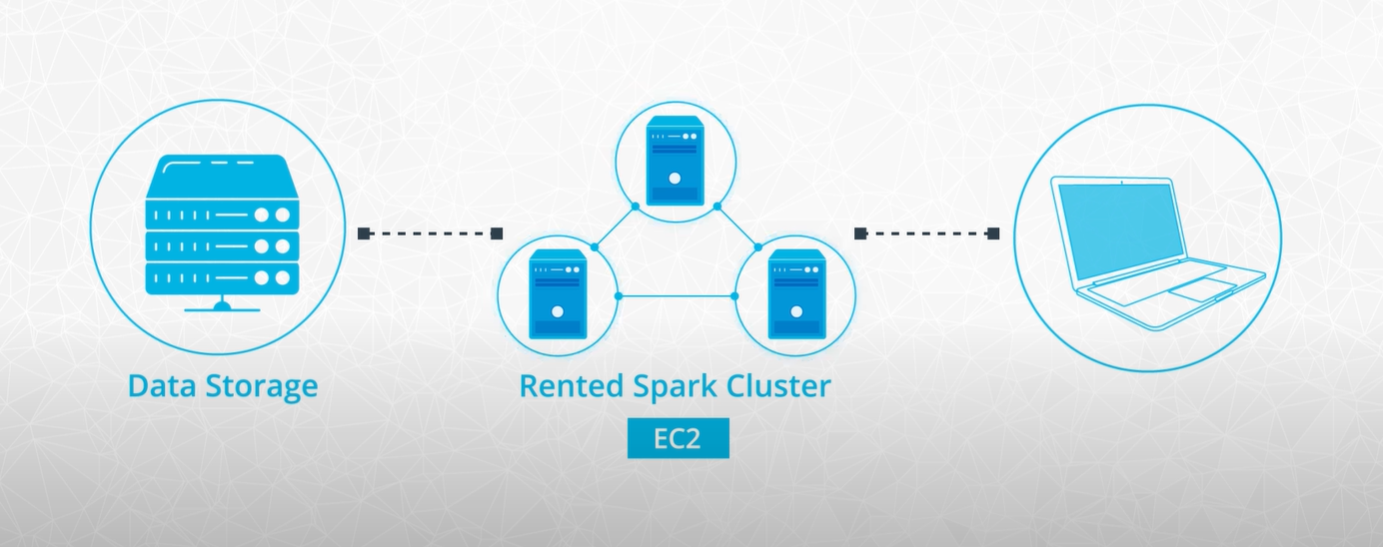
Welcome to the Debugging and Optimization lesson!

In this lesson, you will notice our Instructors primarily teaching through demos and lecture. As debugging and optimization are more conceptual, use this lesson to apply it in the more practical areas of other lessons in the Spark course.

Enjoy!

**Errors in video: (setup instructions AWS)**

1. **AWS does not allow standalone mode**: The video describes that AWS offers to run Spark on Standalone Mode at timestamp 2.50. This is inaccurate. AWS does not allow standalone mode - only YARN mode. More context is available at this [AWS discussion forum post](https://forums.aws.amazon.com/thread.jspa?threadID=231422).
2. **Cost for using AWS EMR clusters:** Please note that for all nodes on the the cluster, AWS charges EMR on top of EC2 cost. The true cost of a cluster with 1 master + 2 core on m5.xlarge (which is more than sufficient for the exercises) would be around $0.75/hour. Note the cost for the EMR clusters on top of AWC EC2 (<https://aws.amazon.com/emr/pricing/)>.  
     
   Remember: **You have limited credits that should be adequate for your exercises. If you forget to shut off any connection with an AWS instance when you are not actively using the AWS instance for the exercise, you may exhaust your credits.**

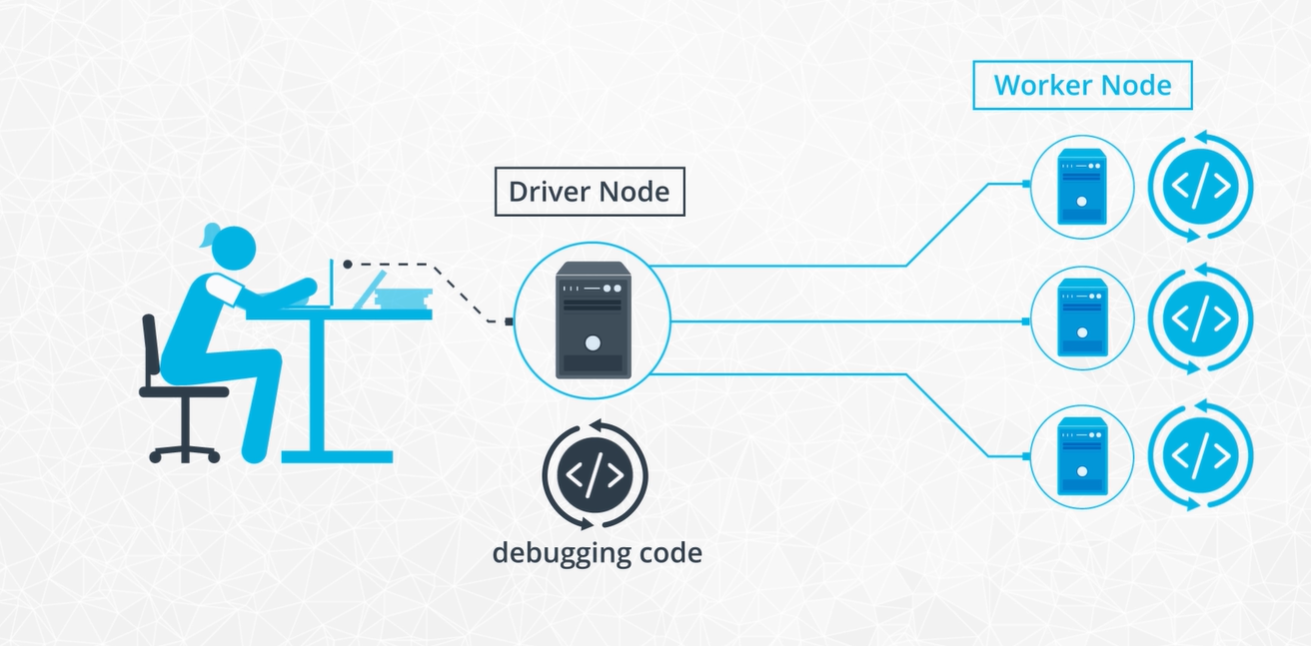


**For Further Optional Reading on the Spark UI**

You may be interested in the [Monitoring and Instrumentation](https://spark.apache.org/docs/latest/monitoring.html) section of the Spark documentation.

### Further Optional Study on Log Data

For further information please see the [Configuring Logging](https://spark.apache.org/docs/latest/configuration.html) section of the Spark documentation.



# Troubleshooting Other Spark Issues

In this lesson, we walked through various examples of Spark issues you can debug based on error messages, loglines and stack traces.

We have also touched on another very common issue with Spark jobs that can be harder to address: everything working fine but just taking a very long time. So what do you do when your Spark job is (too) slow?

## Insufficient resources

Often while there are some possible ways of improvement, processing large data sets just takes a lot longer time than smaller ones even without any big problem in the code or job tuning. Using more resources, either by increasing the number of executors or using more powerful machines, might just not be possible. When you have a slow job it’s useful to understand:

How much data you’re actually processing (compressed file formats can be tricky to interpret) If you can decrease the amount of data to be processed by filtering or aggregating to lower cardinality, And if resource utilization is reasonable.

There are many cases where different stages of a Spark job differ greatly in their resource needs: loading data is typically I/O heavy, some stages might require a lot of memory, others might need a lot of CPU. Understanding these differences might help to optimize the overall performance. Use the Spark UI and logs to collect information on these metrics.

If you run into out of memory errors you might consider increasing the number of partitions. If the memory errors occur over time you can look into why the size of certain objects is increasing too much during the run and if the size can be contained. Also, look for ways of freeing up resources if garbage collection metrics are high.

Certain algorithms (especially ML ones) use the driver to store data the workers share and update during the run. If you see memory issues on the driver check if the algorithm you’re using is pushing too much data there.

## Data skew

If you drill down in the Spark UI to the task level you can see if certain partitions process significantly more data than others and if they are lagging behind. Such symptoms usually indicate a skewed data set. Consider implementing the techniques mentioned in this lesson:

Add an intermediate data processing step with an alternative key Adjust the spark.sql.shuffle.partitions parameter if necessary

The problem with data skew is that it’s very specific to a dataset. You might know ahead of time that certain customers or accounts are expected to generate a lot more activity but the solution for dealing with the skew might strongly depend on how the data looks like. If you need to implement a more general solution (for example for an automated pipeline) it’s recommended to take a more conservative approach (so assume that your data will be skewed) and then monitor how bad the skew really is.

## Inefficient queries

Once your Spark application works it’s worth spending some time to analyze the query it runs. You can use the Spark UI to check the DAG and the jobs and stages it’s built of.

Spark’s query optimizer is called Catalyst. While Catalyst is a powerful tool to turn Python code to an optimized query plan that can run on the JVM it has some limitations when optimizing your code. It will for example push filters in a particular stage as early as possible in the plan but won’t move a filter across stages. It’s your job to make sure that if early filtering is possible without compromising the business logic than you perform this filtering where it’s more appropriate.

It also can’t decide for you how much data you’re shuffling across the cluster. Remember from the first lesson how expensive sending data through the network is. As much as possible try to avoid shuffling unnecessary data. In practice, this means that you need to perform joins and grouped aggregations as late as possible.

When it comes to joins there is more than one strategy to choose from. If one of your data frames are small consider using broadcast hash join instead of a hash join.

## Further reading

Debugging and tuning your Spark application can be a daunting task. There is an ever-growing community out there though, always sharing new ideas and working on improving Spark and its tooling, to make using it easier. So if you have a complicated issue don’t hesitate to reach out to others (via user mailing lists, forums, and Q&A sites).

You can find more information on tuning [Spark](https://spark.apache.org/docs/latest/tuning.html) and [Spark SQL](https://spark.apache.org/docs/latest/sql-performance-tuning.html) in the documentation.