1. Working with GitHub

Git is used for version control while GitHub is a hosting service for Git repositories. It's useful when you're working with a team, but it's still a great tool when you're working by yourself. The lab machines as well as the EC2 instance you're working on have Git installed.

1. As an initial task, let's introduce you to some basic Git commands including git clone, git add, git commit, and git push. After creating an account on GitHub, follow along with this video created by the 722 team.
    a. **Git Clone:** Git clone allows you to clone an already existing repository. Because Git encourages open-source, you have the option to clone code that already exists.
    b. **Git Add, Commit and Push:** You need to type in all three of these commands to push your data from your local repository to GitHub. If you'd like some information on the differences between the three, click here.
2. If you need more information, try going through this tutorial.

2. Fundamentals of Spark and PySpark

What is Spark? Python and R are extremely popular in the context of machine learning but they process data on a single machine. When you have terabytes or petabytes of data this can cause significant delays. This, among other factors, resulted in the creation of Spark which is a cluster-computing framework. Spark allows you to process batch/streaming workloads up to 100x faster than other frameworks such as Hadoop. Spark also allows you to program in a variety of languages including Scala, Java, Python and R.

What is PySpark? PySpark is Spark's Python API. PySpark allows you to use the Python programming language to interact with Spark. This makes interactions with Spark simple and effective.