Couse introduction script

Good morning, good afternoon, good evening and welcome to the course Getting Started With Sparklyr. We’re glad you’re here! My name is Bob Wakefield, and I’ll be your instructor for this course.

I’m rolling into the 18th year of my IT career, and I’ve done everything under the sun IT-wise. I’ve run network cable. I’ve built websites. I’ve done desktop support. I’ve managed applications; you name it, I’ve done it.

About six years ago, I discovered business analytics, and it has since turned into an absolute passion. I got really lucky in life in that I’m basically paid to do my hobby.

It started with crunching big data for a major telecom back before big data wasn’t even a thing. In the process of doing that job, I discovered data science, and from there, data engineering. Early on, I realized that while data science was cool, data engineering was really the market opportunity for me. You can’t do efficient and effective data science without having good, clean data that flows freely through an organization. So I’ve spent the past few years building out data platforms for various organizations and helping executives understand exactly how their business is performing.

<slide>

I am constantly hacking and building things using various tech. When I do something cool or find something cool, I have a tendency to tell everybody I can about it. So if you want to keep up to date on all things big data, feel free to follow me on my various social media outlets.

<slide>

So this course is the first in a series of courses designed to help current data practitioners make the shift from working with small datasets in relational databases to working with gigantic datasets using various newer database technologies. Anybody interested in big data is welcome to take these courses, but they do assume a certain amount of knowledge about data.

We won’t be going to much detail on any one topic. These courses are designed to get you up off the ground quickly and then point you in the right direction for further learning.

In this particular class, we’ll be learning how to work with the Sparklyr package for Apache Spark and the R language. You don’t need any prior knowledge to take this course.

<slide>

This class grew out of a Meet Up presentation I gave on this topic to the Kansas City R Users Group. I ran out of time to cover all the material that I wanted to cover so here we are!

Like I mentioned before, we’re not going to get into too deep of detail on anything. The main goal of this course is to introduce you to some concepts and get you up and running analyzing data with Spark using R quickly.

We’ll talk a little bit about Spark. You don’t need to know Spark in deep detail to work with Sparklyr. But you do need enough basic information so when you are running your R scripts, you know enough to debug any issues.

We’re going to talk about Sparklyr. What it is, what it does, how to get it installed. Later when we get to the labs, we’ll walk through some code samples that demonstrates how Spark and R work together to analyze data.

Lastly, we’ll talk about Databricks. Working on your local machine is great for learning, but eventually, you’re going to want to work on a real Spark Cluster so you can get all the performance benefits that come with working on a cluster.

<slide>

So what is Spark? You first have to understand that Hadoop is old. It’s older than a lot of people think. Hadoop grew out of an academic paper written in the early 2000s. Hadoop was fantastic and solved a lot of problems but we’ve reached a point where it is being surpassed by more advanced technology.

MapReduce was the original framework that we used to process data on a Hadoop cluster. Relatively speaking, MapReduce is slow. Efforts have been made in recent years to speed up processing and that has led to a piece of software called Tez. Of course, if you really want to speed up data processing, the best thing to do is to shove it into memory and work with it there. That’s what Spark lets you do.

Apache Spark is one of the most active Apache projects if not THE most active Apache project out there. It’s got like a billion committers, it’s still being actively developed, and like any popular open source software, it has commercial backing from a company called Databricks.

You can interact with Spark using several different languages and APIs. I’ll be showing you how to work with Spark using the Sparklyr package which leverages the Spark SQL API.

<slide>

So what exactly is the motivation for using Spark with R. I came to it like this. Once I stopped playing with toy datasets and started doing real work, I pretty quickly ran into one of the limitations with R and that’s you are limited to the memory on your local machine.

There are solutions to this. There are some R packages that allow you to swap data to disk but I found these packages difficult to use.

Even if you happen to be working with small data, there are certain algorithms that will explode your dataset beyond your machine’s ability to handle it quickly. Several times I’ve kicked off analysis, went to bed, and woke up to find my machine still crunching away.

Working with your data in Spark, makes a lot of these pains simply go away.

<slide>

There are two R packages for Spark. SparkR and SparklyR. We’re not going to talk to much about Sparkr, I just wanted you to be aware that it exist. Sparkr works with core R. At the time I was learning all this stuff, I was also learning about the Tidy Verse. The Tidy Verse, if you’re not aware is an ecosystem of packages started by Hadley Wickam to more easily deal with data. One of those packages is dplyr. Sparklyr works WITH the Tidy Verse. Specifically it allows you to work with your data the same as working with it using dplyr. More on this later.

Suffice to say, I was drawn to Sparklyr because of its interoperability with the Tidyverse and didn’t bother to spend any time learning Sparkr.

Sparklyr IS a product. It’s free open source but it was created by the people that make R Studio. That used to be a good and bad thing. The good thing is, there is a lot of integration with R studio. But It used to be the case that, if you wanted to work with big data in a real cluster, you had to use a product called R Studio Server which was a paid product so that was a downside. However, you now have options. You can use Databricks as we’ll see later, or you can use the free open source version of R Studio Server which wasn’t free when I started learning Sparklyr in the summer of 2017.

In this course, I’ll show you how to work with sparklr on your local machine and in the cloud with Databricks, so we won’t need to worry about R Studio Server.

<slide>

Working with Sparklyr is a little weird. It’s not quite like working with normal R. There are some quirks you need to be aware of.

The main thing is reasoning about where exactly your data is. When you work with Sparklyr, data lives in two places. On the cluster, and on your local machine.

A normal workflow sounds something like this. You put data on your cluster. You connect to your cluster using Sparklr. You create predictive models ON the cluster. You collect the results of your analysis and bring it back to your local machine. The results are much smaller than the data you were working with. Once you have results, you can do things like plot it using ggplot2 on your local machine. We’ll see this workflow later in the lab section of this course.

While you are working with your data, R Studio will be keeping track of what objects have been created where. It will be up to you to understand which objects are actual data and what objects are actually pointers to data.

<slide>

So up to this point, I’ve been hitting you with a lot of information really fast with the intention of getting into deeper detail later. Let’s take a moment and talk about one of the concepts I mentioned earlier.

I got introduced to Spark many many years ago but I never really understood it. The website seemed impenetrable. I didn’t really understand what it was and what it can do. At the top of 2017, I noticed that they had made some changes to the software and for some reason, a light bulb went off in my head. It was much easier to understand what Spark was all about and I made a concerted effort to learn it.

As I mentioned before, there all kinds of ways to interact with Spark. Of course, this wasn’t always the case. The original way to work with data in Spark was a bit rough to get your head around and that was a major barrier for me.

I’ve been working with data for the better part of two decades. In all this time, I’ve been working with data pretty much the same way; a two-dimensional rectangular dataset. You have rows and columns and getting to data was a matter of locating on a XY coordinate on a cartesian plane. Early Spark didn’t work like that.

Spark had something called an RDD. A resilient distributed data set. It was great for working with unstructured data but it just wasn’t that rectangular dataset that I was used to seeing. SparkSQL, in conjunction with the dataset/dataframe API changed all that. Now you could load data into Spark and reason about it just like you would any other data set and, as an extra added bonus, you could leverage your existing SQL skills to access the data.

All of that has made Spark a lot more accessible to the average analyst.

<slide>

It should go without saying but you need to understand that Spark is its own separate thing that doesn’t come with R or R Studio. It needs to be installed. In lab, I’ll show you how to install it on your local machine. Installing on a cluster is a little harder and beyond the scope of this course.

If you want to work on a cluster, there are several different options to connect to the cluster that all depend on how you have the cluster set up. For my money, I skip all the hard work of setting up a cluster and just use Databricks.

<slide>

So what exactly IS Databricks? It’s a Spark platform as a service in the cloud. In the labs, I’ll show you how to work with Databricks. It’s a really great way to learn Spark. It’s really easy to learn and understand especially if you’re already used to working with notebooks like Jupyter.

I’ll be showing you how to work the free community version of Databricks. The free version has limited space but it’s enough that you can use it with most Kaggle Competition datasets that I run across.

<slide>

So that’s Sparklyr in a nutshell. In the labs, we’ll go over some basic examples. It will be enough information to get you on your way to analyzing big data with Spark and R. See you in the labs!