Getting Started With Sparklyr

Bob Wakefield
Principal
bob@MassStreet.net
Twitter:
@BobLovesData

Mass Street
Analytics

Bob’s Background

- IT professional 17 years
- Currently working as a Data Engineer
- Education
  - BS Business Admin (MIS) from KState
  - MBA (finance concentration) from KU
  - Coursework in Mathematics at Washburn
  - Graduate certificate Data Science from Rockhurst
- Addicted to everything data

Follow Me!

- Personal Twitter: @BobLovesData
- Company Twitter: @MassStreet
- Blog: DataDrivenPerspectives.com
- Website: www.MassStreet.net
- Facebook: @MassStreetAnalyticsLLC

About Mass Street's Hands-On Series Of Courses

- Designed to introduce students to various data management concepts
- All classes are introductory in nature and do not do "deep dives" into any one topic
- Every course has labs to reinforce learning
- All classes end with recommendations on how to continue your learning
- Anybody is welcome to take the course
- Specifically built for data professionals who want to get into Big Data

Mass Street Analytics

What We’ll Cover

- We’re gonna hit stuff at 10,000 ft.
- Spark
  - General information
- Sparklyr
  - How to get everything installed locally
  - Walk through some sample code
- Databricks
  - How to do Sparkr on a cluster

What Is Spark

- Distributed in memory processing framework
- Rapidly replacing MapReduce as a means to crunch data.
- Many ways to interact with Spark
  - R, Java, Scala, Python
  - Sparkr, Sparklyr, H2O with Sparkling Water
- Several APIs
  - RDDs, Dataset/Dataframe, Spark Streaming, Spark SQL, Spark Streaming, Strucutred Streaming

Why would you want to use Spark?

- To get around memory limitations in R.

Mass Street Analytics

What is Sparklyr?

- There are two R packages for Spark
  - Sparkr and Sparklyr
- Sparklyr is a product of the folks that make R Studio
  - That used to mean strings attached
  - R Studio Server has been open sourced

What is Sparklyr?

- Sparklyr allows you to work with data on a spark cluster using dplyr.
- Uses the SparkSQL API
- Little bit weird. Not like working with normal R.

What is SparkSQL?

- The new way to interact with Spark.
- Another SQL on big data implementation.
- Allows you to write SQL statements against a spark cluster.
- You can use straight SQL or dplyr techniques
- All super easy with R.

Cluster Connection Methods

- In every case you need a copy of Spark
    - Locally
    - On a cluster
- Options for connecting to a cluster
    - R Studio Server
        - Mesos/Yarn
        - Spark Standalone
    - Livy
    - Just use Databricks

What is Databricks?

- Spark P/SaaS
- Basically a worry free Spark cluster
- Two versions
  - Commercial
  - Community
- Community version
  - Practice on a real cluster
  - Limited on space

Examples