

Introduction to Spark with R

CDC - RStudio PBC

Agenda

- Spark Overview
- Spark with R
- Spark Clusters
- Resources

Spark Overview



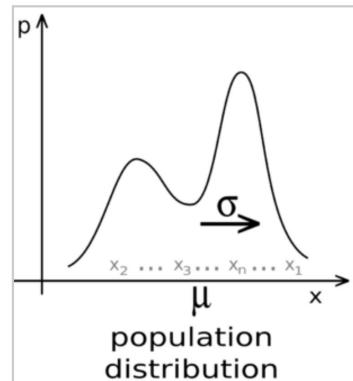
What to do when code is slow?

```
lm(mpg ~ wt + cyl, mtcars)
```

What to do when code is slow?

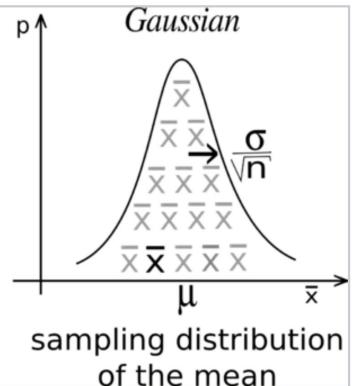
Sample

```
mtcars %>% dplyr::sample_n(10) %>% lm(mpg ~ wt + cyl, .)
```



samples
of size n

A diagram showing a horizontal arrow with two points labeled \bar{x} and \bar{x} above it, representing samples of size n drawn from the population distribution.



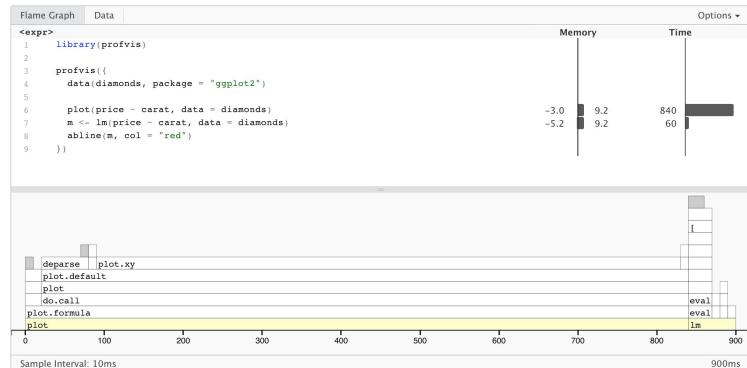
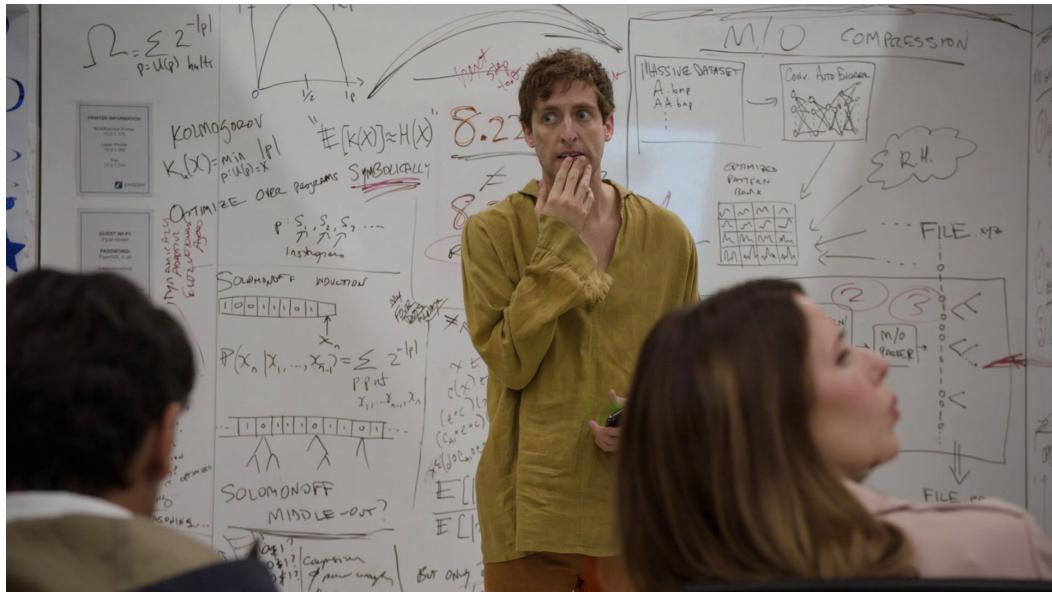
Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem.^[2]



What to do when code is slow?

Profile

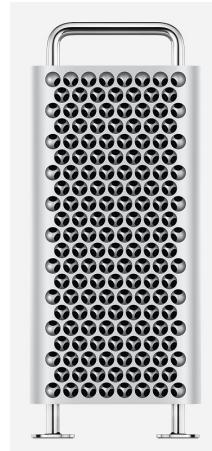
```
profvis::profvis(mtcars %>% lm(mpg ~ wt + cyl, .))
```



What to do when code is slow?

Scale Up

```
renv::snapshot()  
cloudml::cloudml_train("train.R")  
aws.ec2::run_instances()
```



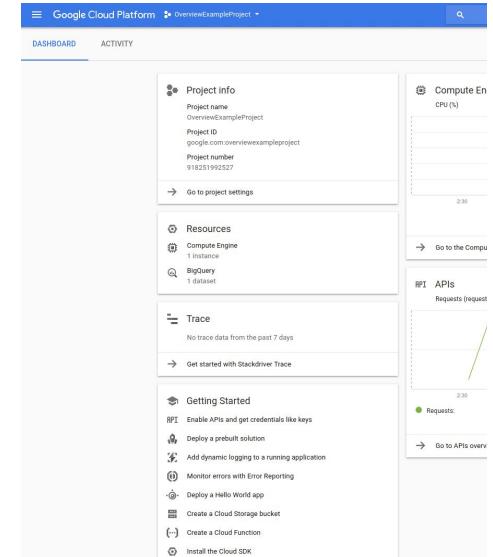
Screenshot of the AWS CloudFormation 'Step 1: Choose an Amazon Machine Image (AMI)' interface. The interface shows a list of available AMIs:

- Red Hat Enterprise Linux 6.4 - ami-a25415cb (64-bit) / ami-7e175617 (32-bit)
Free tier eligible
- SUSE Linux Enterprise Server 11 - ami-e8084981 (64-bit) / ami-b60948df (32-bit)
Free tier eligible
- Ubuntu Server 12.04.3 LTS - ami-a73264ce (64-bit) / ami-a53264cc (32-bit)
Free tier eligible
- Ubuntu Server 13.10 - ami-ad184ac4 (64-bit) / ami-a9184ac0 (32-bit)
Free tier eligible
- Amazon Linux AMI (HVM) 2013.09 - ami-08792c00
The Amazon Linux AMI is an EBS-backed, HVM image. It includes Linux 3.4, AWS tools, and Tomcat.
Root device type: ebs Virtualization type: hvm
- Red Hat Enterprise Linux 6.4 for Cluster Instances - ami-321805b6

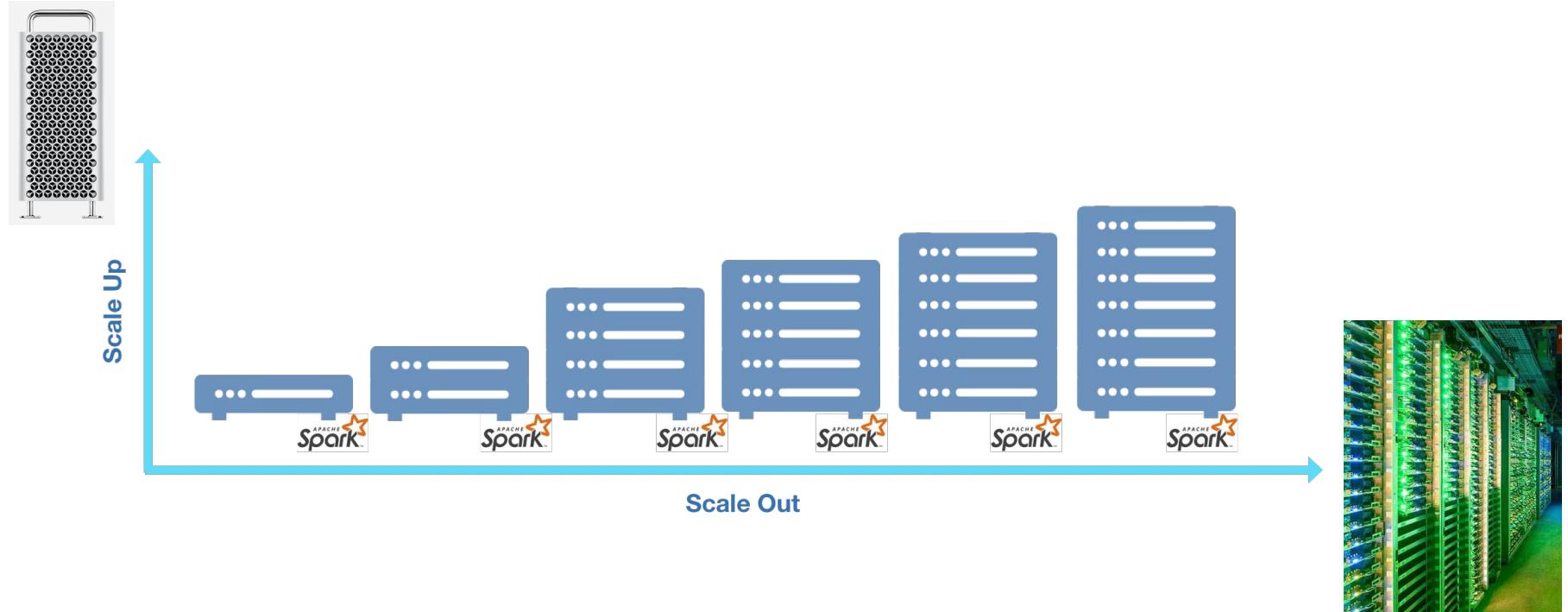
What to do when code is slow?

Scale Out

```
mtcars_tbl %>% sparklyr::ml_linear_regression(mpg ~ wt + cyl)
```

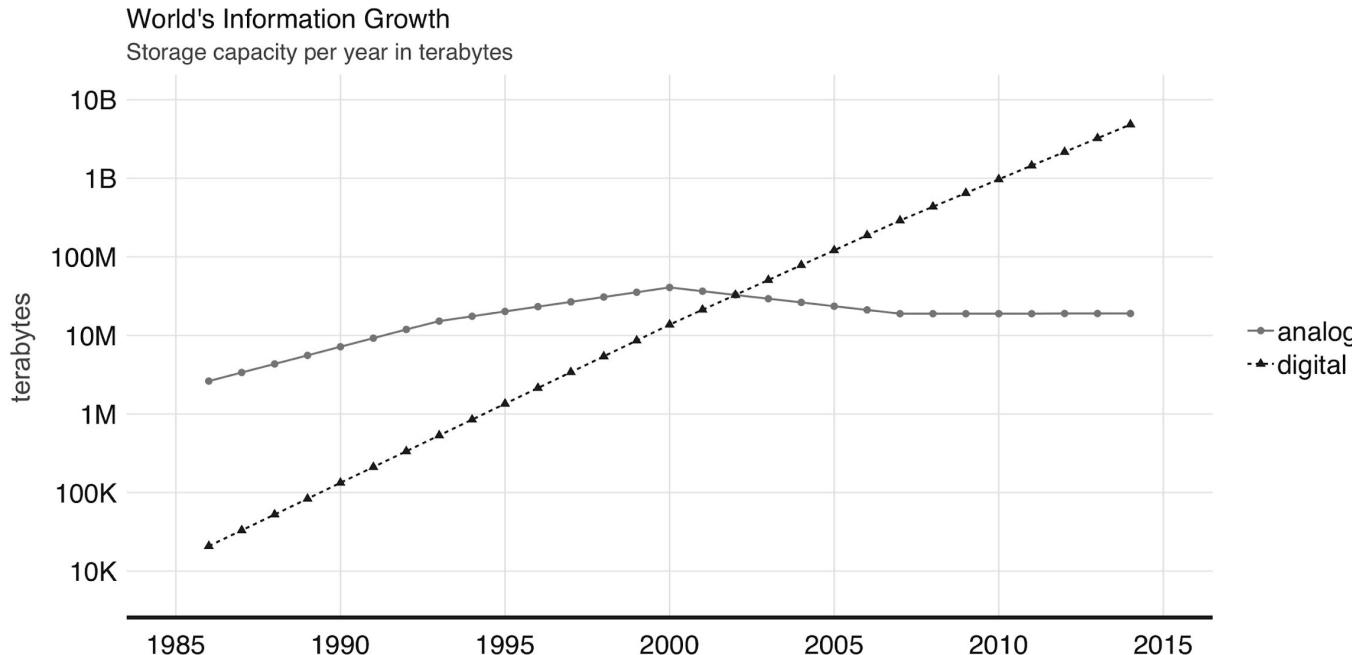


Scaling Out with R and Spark



Why is Spark important?

World Bank report shows information growing exponentially.



O'REILLY®

**Mastering
Spark with R**

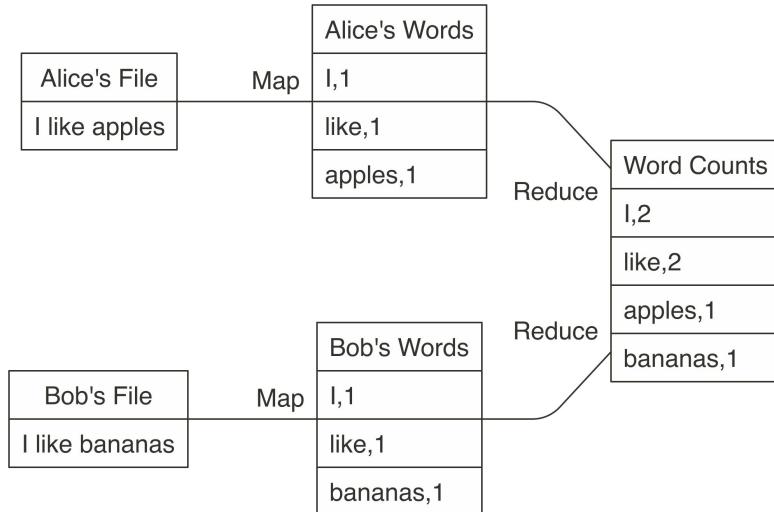
The Complete Guide to Large-Scale Analysis
and Modeling



Javier Luraschi,
Kevin Kuo & Edgar Ruiz
Foreword by Matei Zaharia

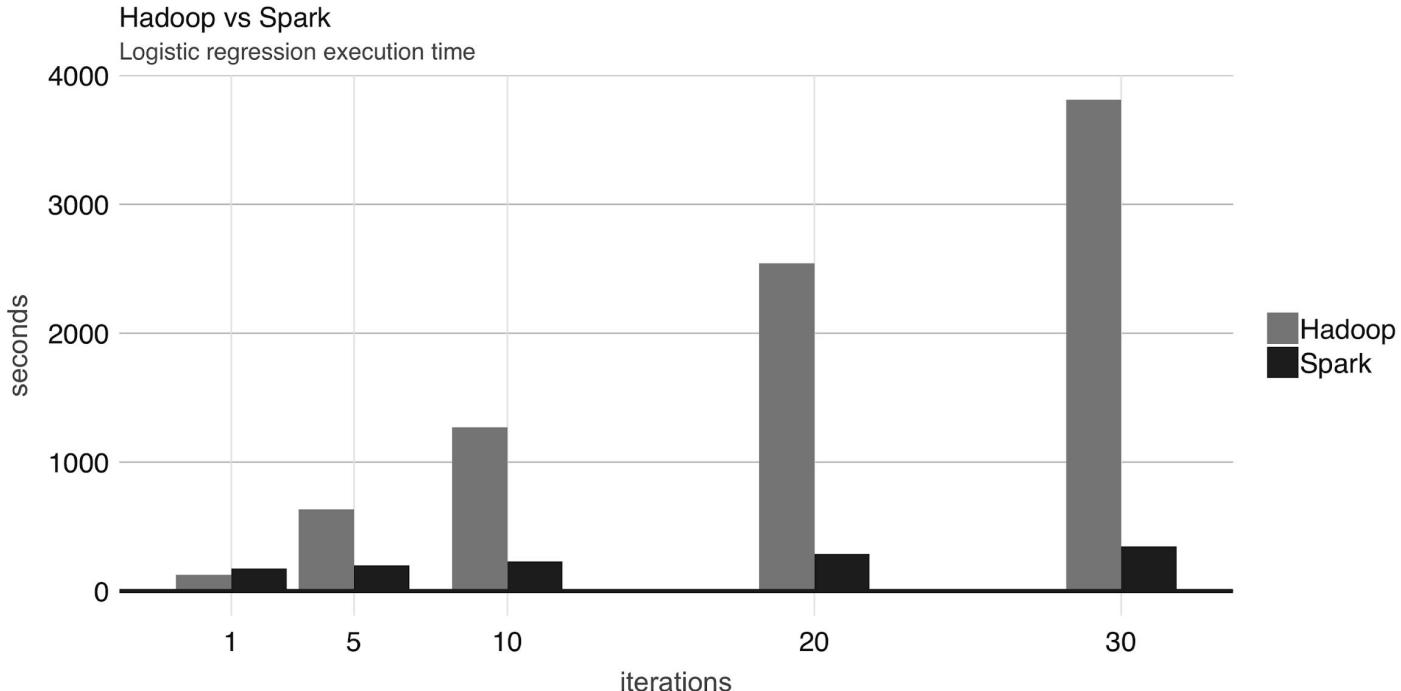
What is Hadoop and MapReduce?

Google develops a distributed file systems and MapReduce, which Yahoo open sources into Hadoop.



Why do we need Spark?

Berkeley improves Hadoop with the Apache Spark project.



What is Spark?

From spark.apache.org, Spark is: easy to use, general, fast and runs everywhere.

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including [SQL and DataFrames](#), [MLlib](#) for machine learning, [GraphX](#), and [Spark Streaming](#). You can combine these libraries seamlessly in the same application.



Runs Everywhere

Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.

You can run Spark using its [standalone cluster mode](#), on [EC2](#), on [Hadoop YARN](#), on [Mesos](#), or on [Kubernetes](#). Access data in [HDFS](#), [Alluxio](#), [Apache Cassandra](#), [Apache HBase](#), [Apache Hive](#), and hundreds of other data sources.



Spark with R

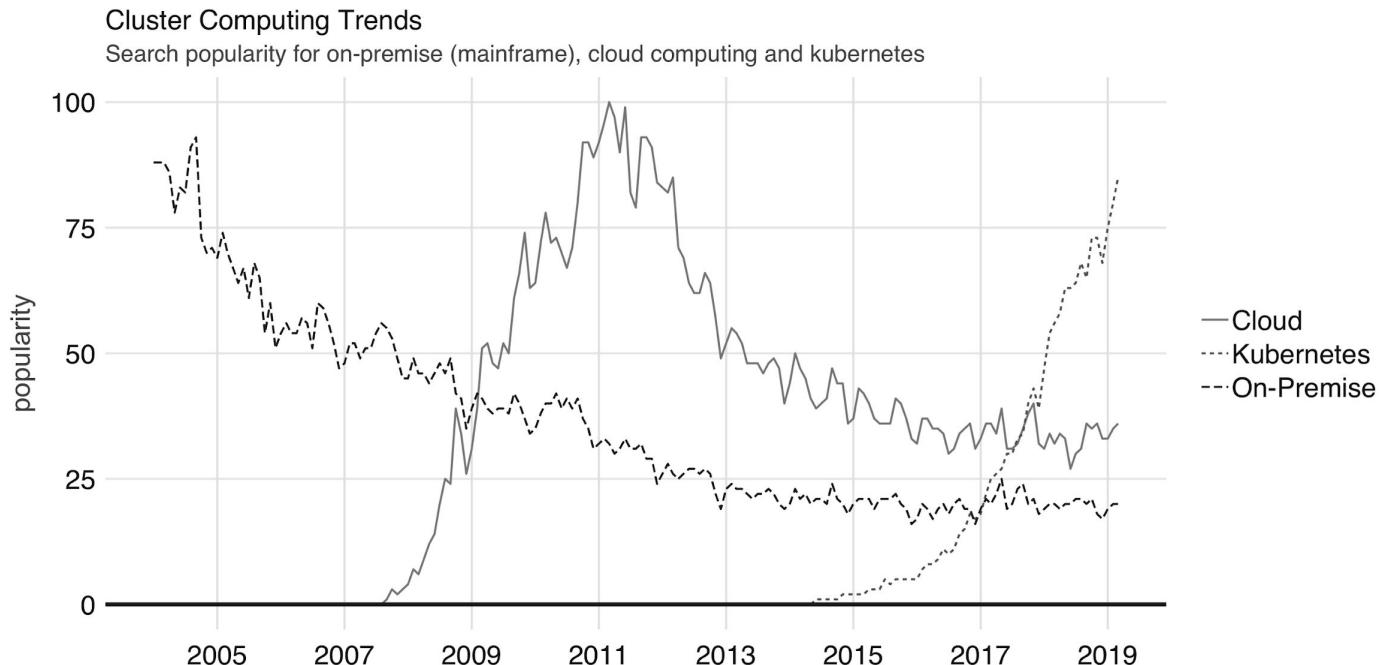
Using Spark in R with sparklyr

```
spark_install()  
sc <- spark_connect(master = "local")  
  
cars <- spark_read_csv(sc, "cars", "input/")  
  
summarize(cars, n = n())  
dbGetQuery(sc, "SELECT count(*) FROM cars")  
  
ml_linear_regression(cars, mpg ~ wt + cyl)  
  
ml_pipeline(sc) %>%  
  ft_r_formula(mpg ~ wt + cyl) %>%  
  ml_linear_regression()  
  
spark_context(sc) %>% invoke("version")  
spark_apply(cars, nrow)  
  
stream_read_csv(sc, "input/") %>%  
  filter(mpg > 30) %>%  
  stream_write_json("output/")  
  
# Install local Spark  
# Connect to Spark cluster  
  
# Read data in Spark  
  
# Count records with dplyr  
# Count records with DBI  
  
# Perform linear regression  
  
# Define Spark pipeline  
# Add formula transformation  
# Add model to pipeline  
  
# Extend sparklyr with Scala  
# Extend sparklyr with R  
  
# Define Spark stream  
# Add dplyr transformation  
# Start processing stream
```

Spark Clusters

Where do I get a Spark cluster from?

On-premise (you own the computers), cloud (you rent them), or Kubernetes (you don't care).



Spark Clusters in the Cloud

You can create a cluster from Amazon, Databricks, Google, Microsoft, IBM, Qubole, etc.

```
sparklyr-0.9.r
Attached:  sparklyr  File  View: Code  Permissions  Run All  Clear  Publish  Comments  Revision history

  Home  Help  Workspaces  Recent  Data  Cluster  Jobs  Search

  1: install.packages("sparklyr")
  2: library(sparklyr)

Installing package into '/data/trcks/spark/R/lib'
(as 'lib' is unspecified)
 打造成一个可重用的依赖项，`sparklyr`，`httr`，`htmlwidgets`，`httpuv`，`xtable`，`rsrcools`，`later`，`promise`，`config`，`dplyr`，`r2dft`，`rsparklyr`，`shiny`，`Forge`和`curl`。
  Command took 0.93 minutes -- by javier@studio.com at 11/28/2018, 11:09:04 AM on sparklyr
  3: 
  4: sc <- spark_connect(method = "datastricks")

Command took 1.08 seconds -- by javier@studio.com at 11/28/2018, 11:13:07 AM on sparklyr

Shift+Enter to run  sparklyr
```

The screenshot shows the Azure portal interface for creating a new service principal. The top navigation bar includes 'All services' (dropdown), 'Search', 'Create resource', 'My projects', 'Dashboard', 'Logs', 'Metrics', 'Feedback', and 'Help & support'. The left sidebar has sections for 'My projects' (dropdown), 'Dashboard', 'Logs', 'Metrics', 'Feedback', and 'Help & support'. The main content area has tabs for 'Overview', 'Identity', 'API permissions', 'Certificates & secrets', 'Service principals', 'Role assignments', and 'Audit logs'. A search bar at the top right contains 'Search resources' and a dropdown for 'Region'. The 'Identity' tab is selected.

Service principal details

Name: GitHub App
Description: GitHub App - https://github.com/.../github-app

Type: OAuth Client-credentials

Resource names: GitHub App

Grant types: OAuth 2.0 Client Credentials

Access tokens

Associated services

Service principal details

Name: GitHub App
Service type: GitHub App

Access tokens

Role assignments

Identity

Metrics

Audit logs

Feedback

Help & support

The screenshot shows the Quoter interface for managing clusters. At the top, there's a navigation bar with 'Clusters' selected, followed by 'Clusters' and 'Account: new_package_management'. A search bar and various icons are also present. Below the header, a 'New' button is visible. The main area is titled 'Active Clusters' and lists one cluster entry:

	29816			2 Total (0 Spot, 0 Spot Block)	12 Sep 2019 19:12, ameya@quoter.com
--	--------------	--	--	--------------------------------	-------------------------------------

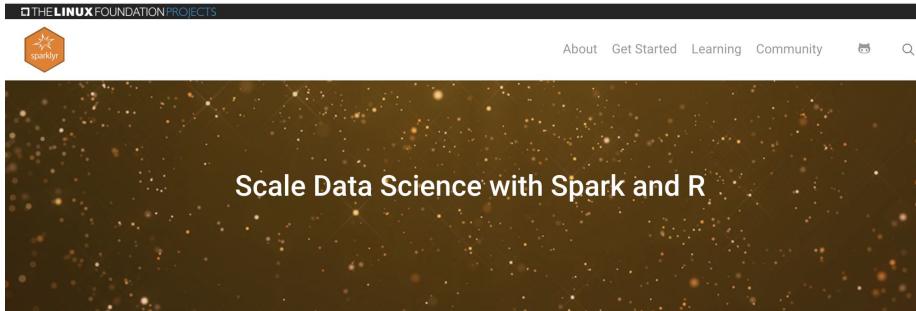
Below this, a section for 'View Deleted Clusters' is shown. On the right side, a 'Resource Manager' panel is open, displaying options like 'Stop', 'Edit', and 'Copy Master DNS'. The 'Resource Manager' section itself includes links for 'DFS Status', 'AutoScaling Logs', 'Jupyter', 'Spark History Server', 'RStudio', 'Cluster Usage', and 'Cluster Start Logs'.



Resources

Linux Foundation

Sparklyr became a Linux Foundation project in 2020, read more on sparklyr.ai



About

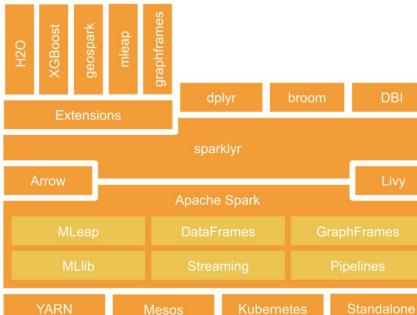
sparklyr is an open-source and modern interface to scale data science and machine learning workflows using **Apache Spark™**, R, and a rich **extension ecosystem**.

It enables using Apache Spark with ease using R by providing access to core functionality like installing, connecting and managing Spark and using Spark's **Mlib**, Spark **Structured Streaming** and Spark **Pipelines** from R.

Supports well-known R packages like **dplyr**, **DBI** and **broom** to reduce the cognitive overhead from having to re-learn libraries.

And enables a rich-ecosystem of extensions to use in Spark and R: **XGBoost**, **MLeap**, **GraphFrames**, **H2O**, and optionally enable **Apache Arrow** to significantly improve performance.

Through Spark, this allows you to scale your Data Science workflows in **Hadoop YARN**, **Mesos**, **Kubernetes** or **Apache Livy**.



Sponsors and Users

Support across all major platforms, various supporters and many users.



EXPRESS SCRIPTS®

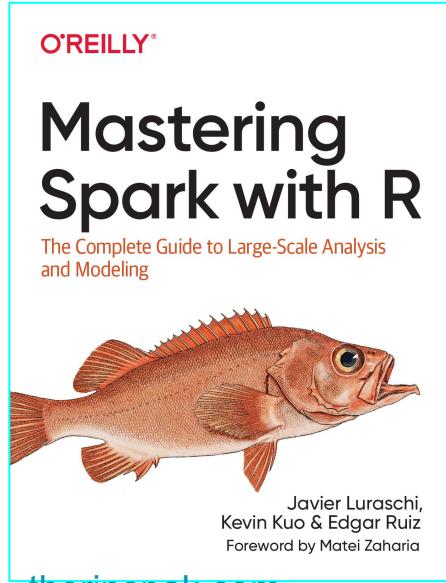


Ketchbrook Analytics



Books, Blogs and Videos

Thanks! - javier@rstudio.com - @javierluraschi



therinspak.com

blogs.rstudio.com/ai

youtube.com/c/mlverse