Introduction to Spark with R

CDC - RStudio PBC

Agenda

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

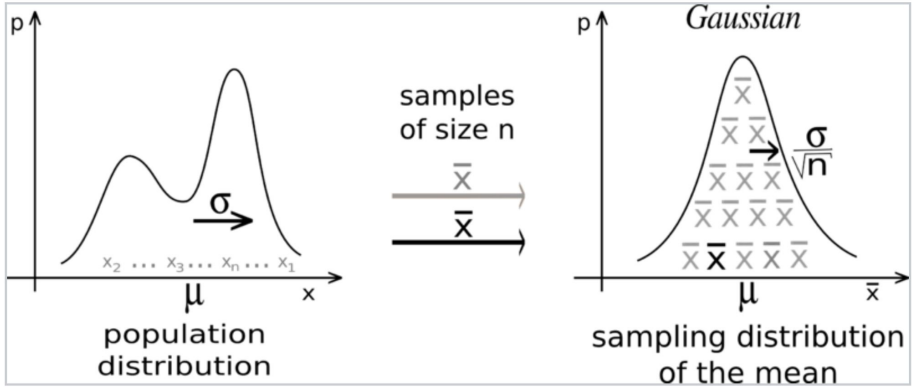
Spark Overview

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, .)
```

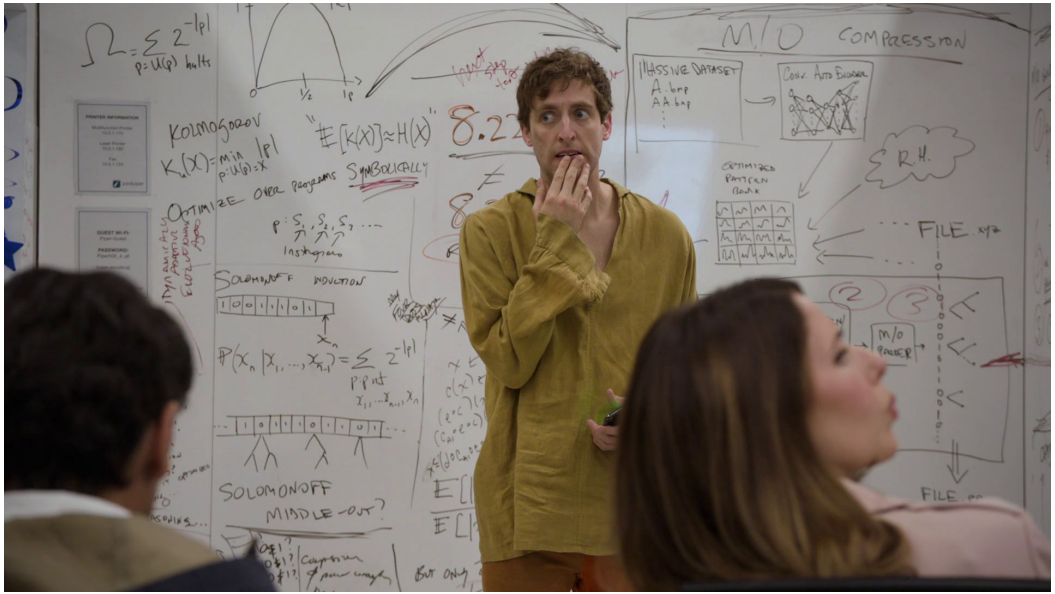


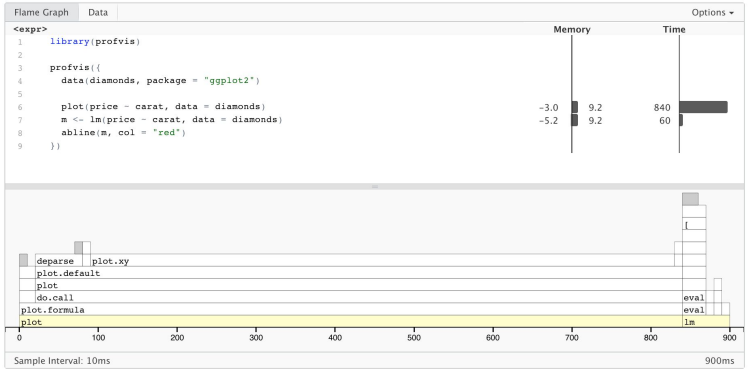
Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem.[2]

What to do when code is slow?

 # Profile

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

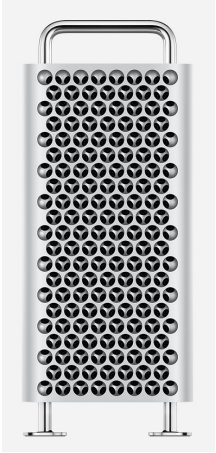


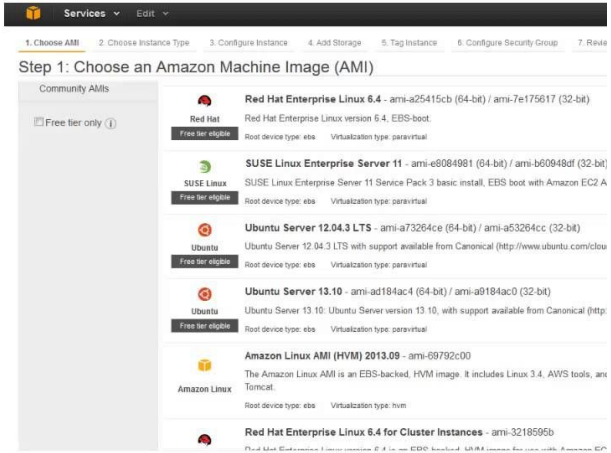


What to do when code is slow?

 # Scale Up

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

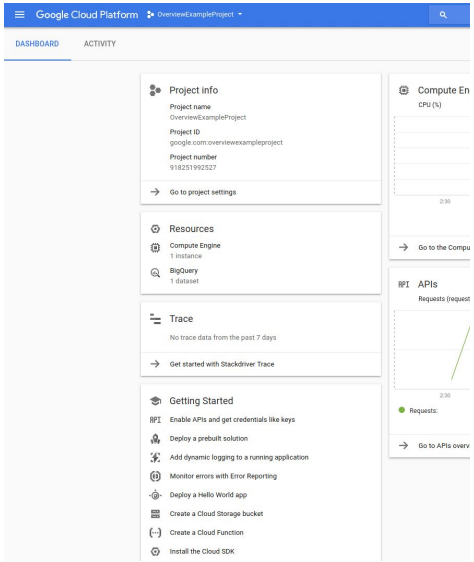




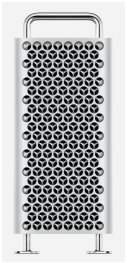
What to do when code is slow?

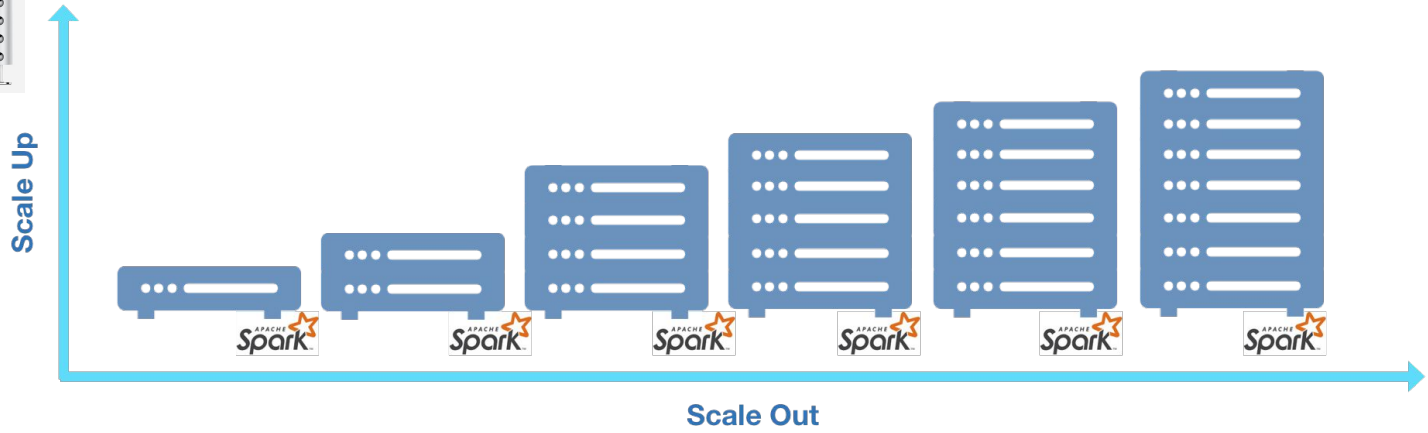
 # Scale Out

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



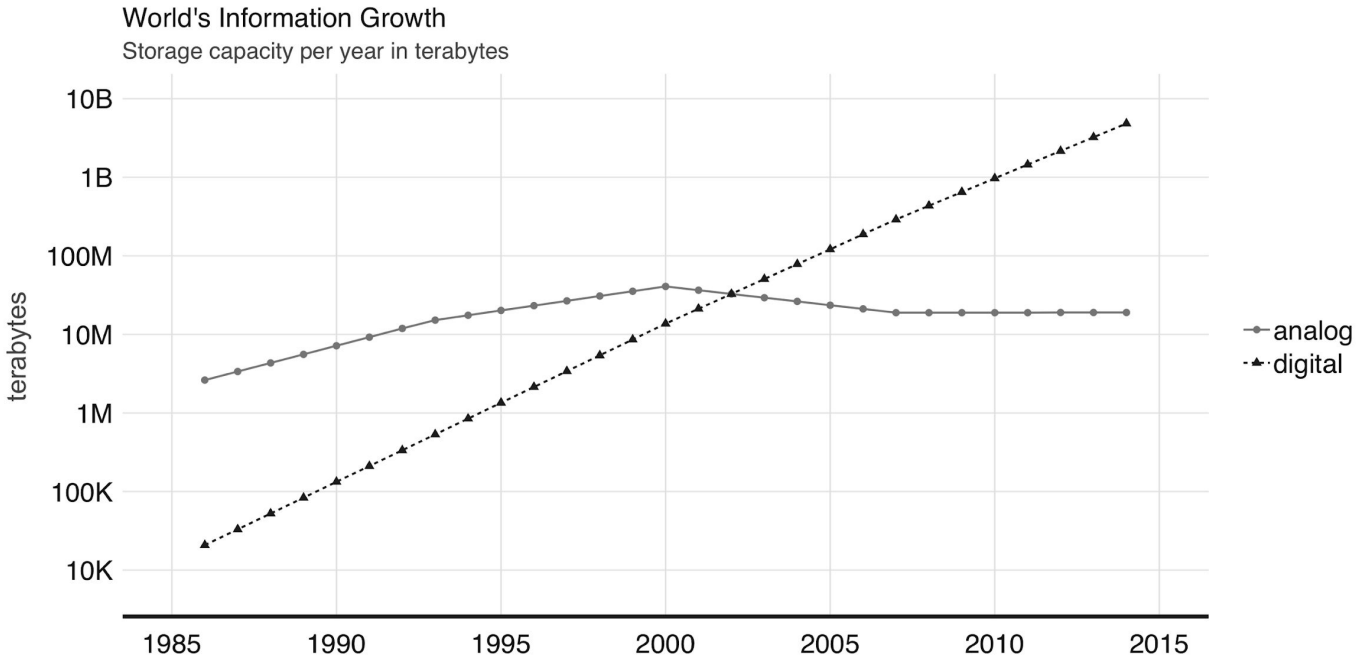
Scaling Out with R and Spark

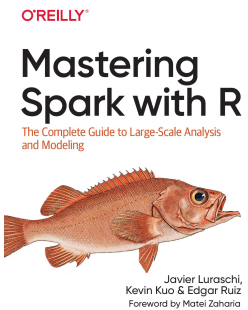




Why is Spark important?

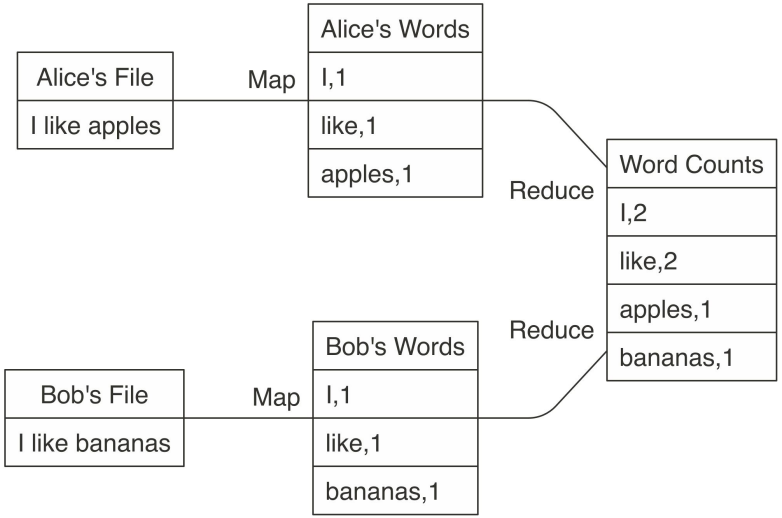
World Bank report shows information growing exponentially.





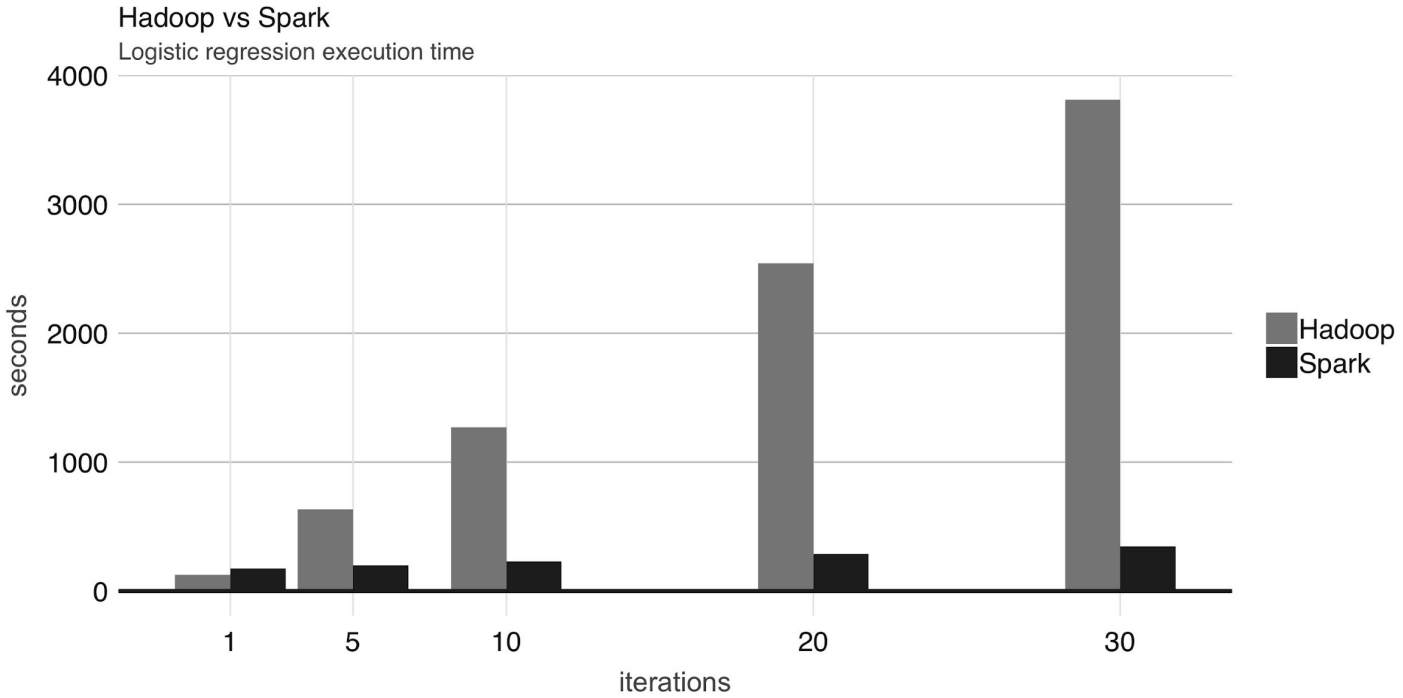
What is Hadoop and MapReduce?

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

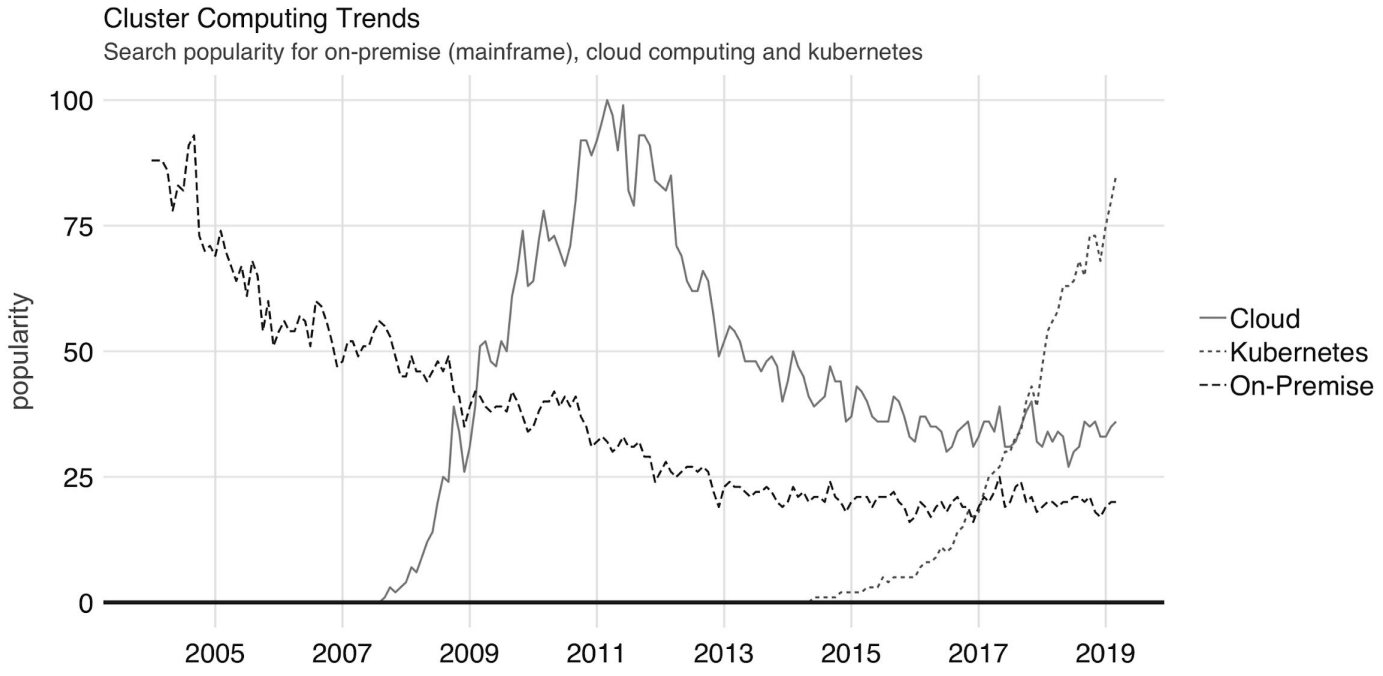
Using Spark in R with sparklyr

```
spark_install()                                          # Install local Spark
sc <- spark_connect(master = "local")                    # Connect to Spark cluster

cars <- spark_read_csv(sc, "cars", "input/")             # Read data in Spark

summarize(cars, n = n())                                 # Count records with dplyr
dbGetQuery(sc, "SELECT count(*) FROM cars")              # Count records with DBI

ml_linear_regression(cars, mpg ~ wt + cyl)               # Perform linear regression

ml_pipeline(sc) %>%                                      # Define Spark pipeline
  ft_r_formula(mpg ~ wt + cyl) %>%                       # Add formula transformation
  ml_linear_regression()                                 # Add model to pipeline

spark_context(sc) %>% invoke("version")                  # Extend sparklyr with Scala
spark_apply(cars, nrow)                                  # Extend sparklyr with R

stream_read_csv(sc, "input/") %>%                        # Define Spark stream
  filter(mpg > 30) %>%                                   # Add dplyr transformation
  stream_write_json("output/")                           # Start processing stream
```

Spark Clusters

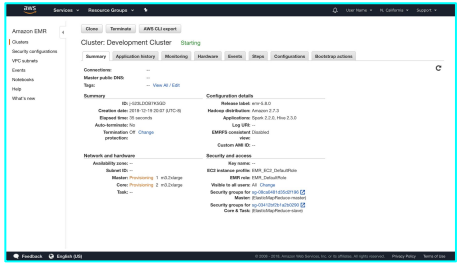
Where do I get a Spark cluster from?

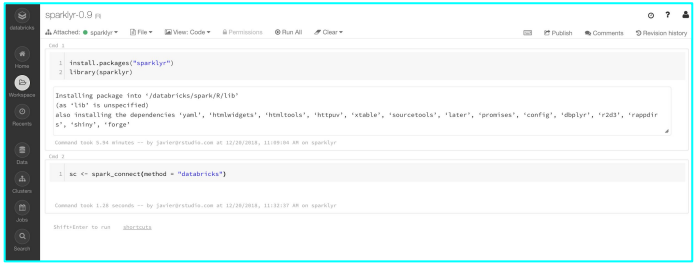
On-premise (you own the computers), cloud (you rent them), or Kubernetes (you don’t care).

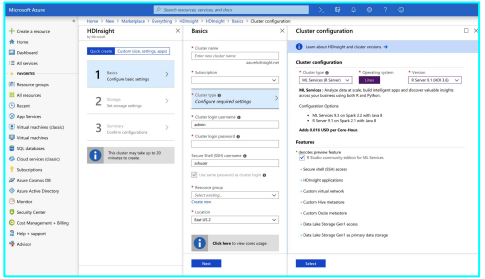


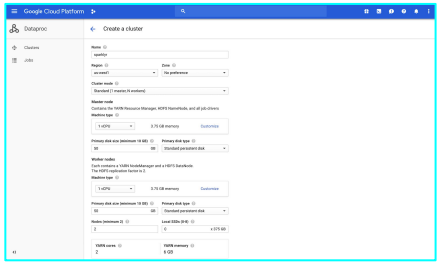
Spark Clusters in the Cloud

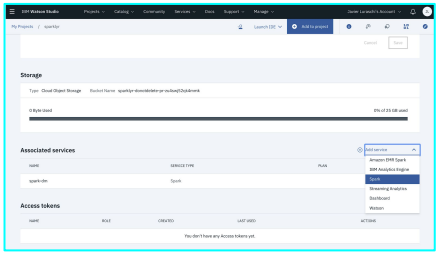
You can create a cluster from Amazon, Databricks, Google, Microsoft, IBM, Qubole, etc.

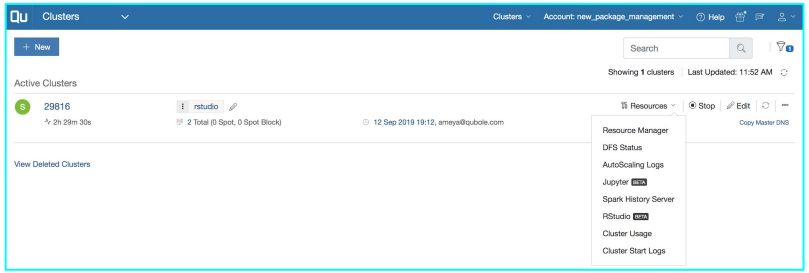








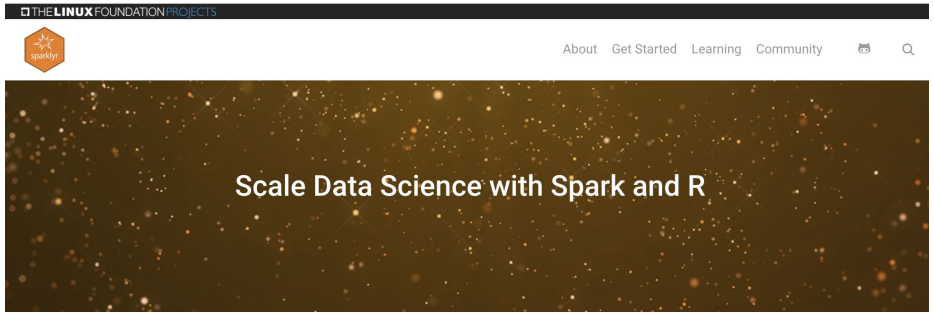




Resources

Linux Foundation

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

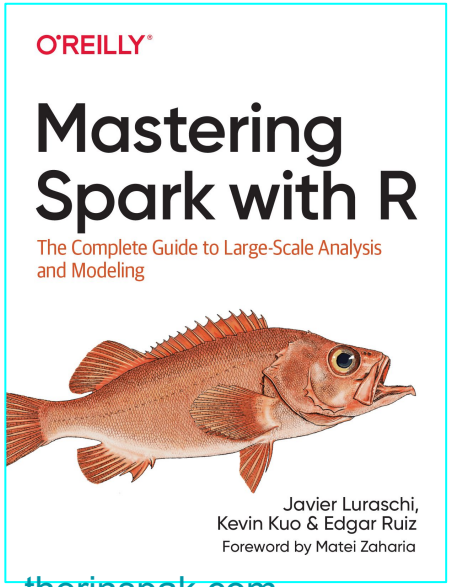


Sponsors and Users

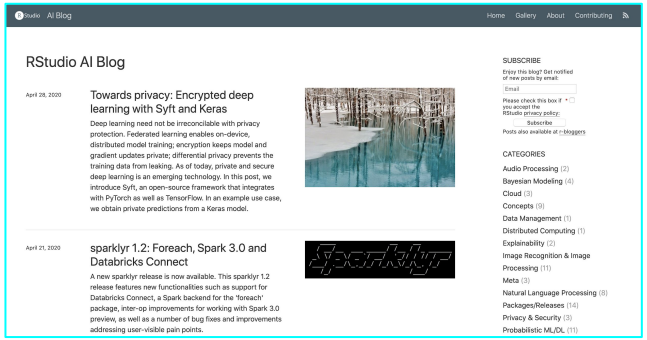
Support across all major platforms, various supporters and many users.

Books, Blogs and Videos

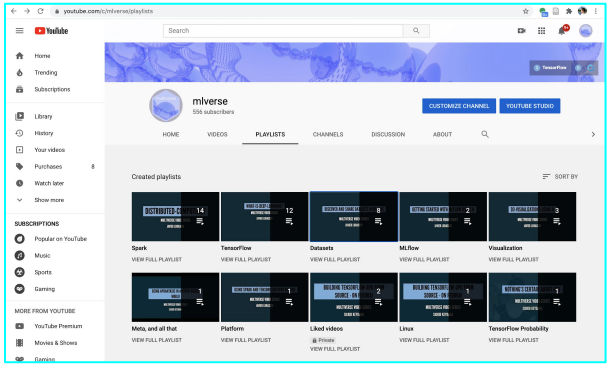
Thanks! - javier@rstudio.com - @javierluraschi



therinspak.com



blogs.rstudio.com/ai



youtube.com/c/mlverse