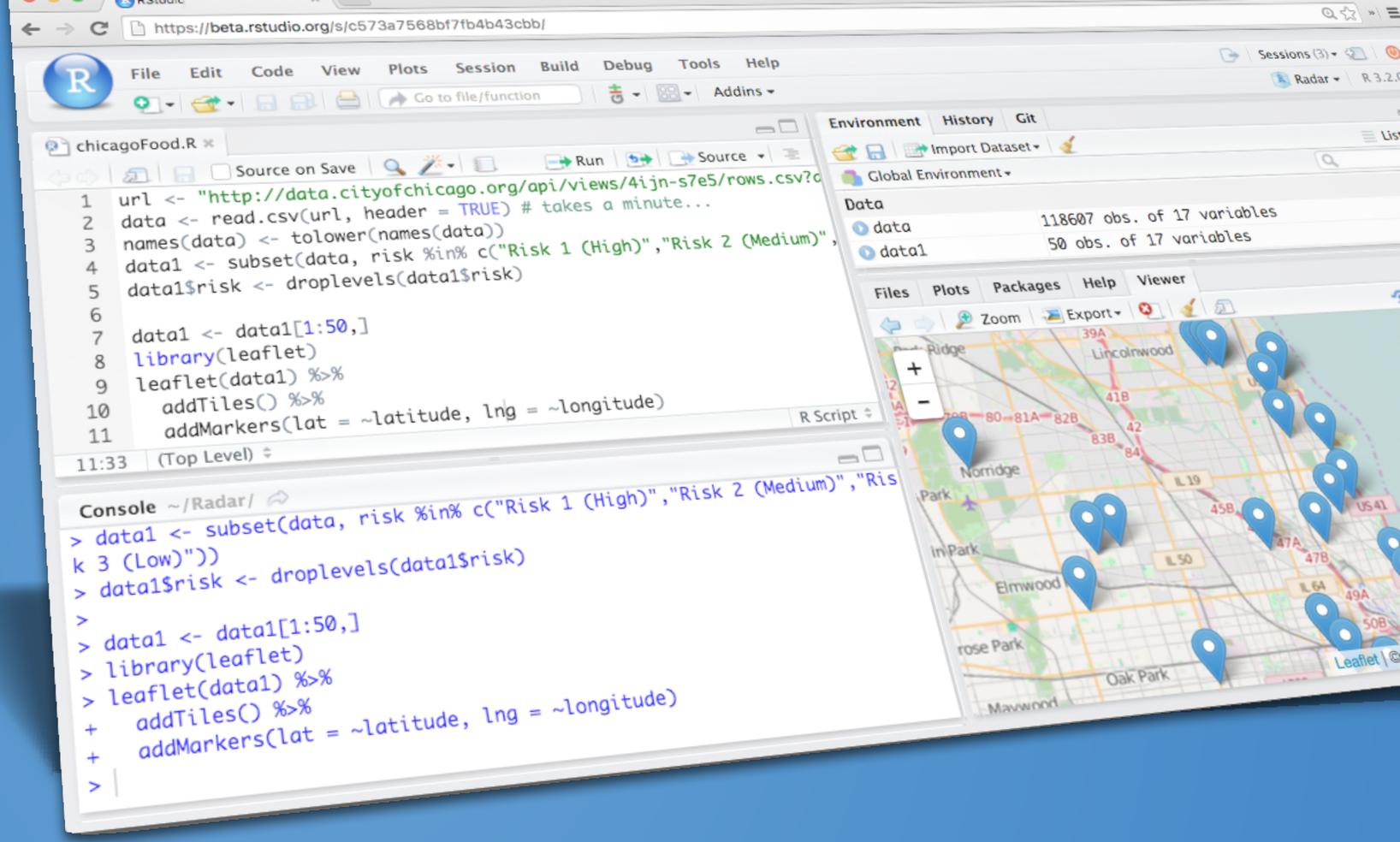
RStudio®



sparklyr

Understanding Spark and sparklyr deployment modes

SPARKLYR - AUTHORS

Javier Luraschi – author , creator

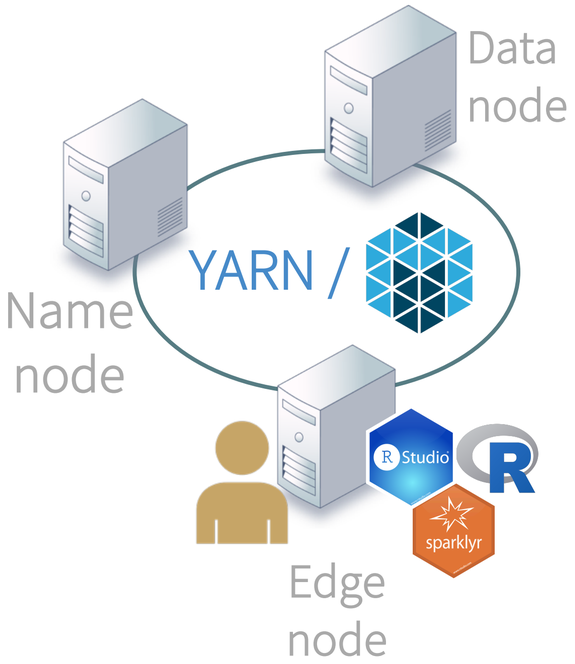
JJ Allaire – author

Kevin Ushey – author

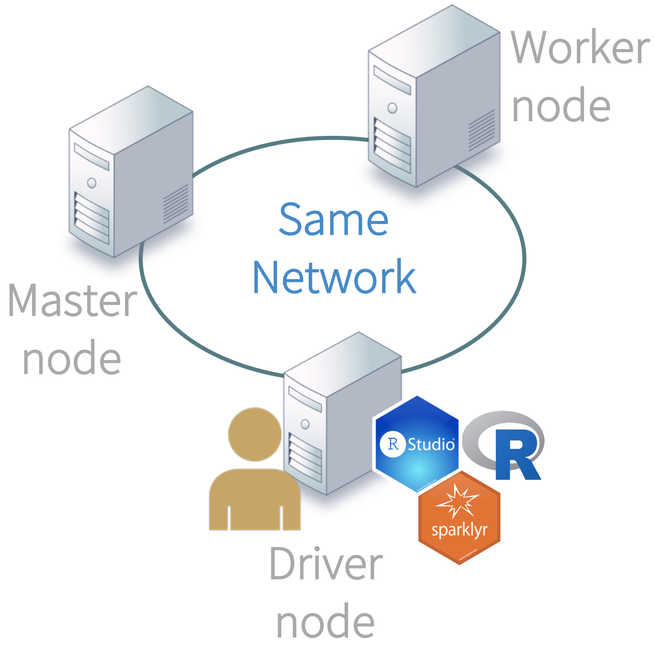
RStudio®

DEPLOYMENT MODES

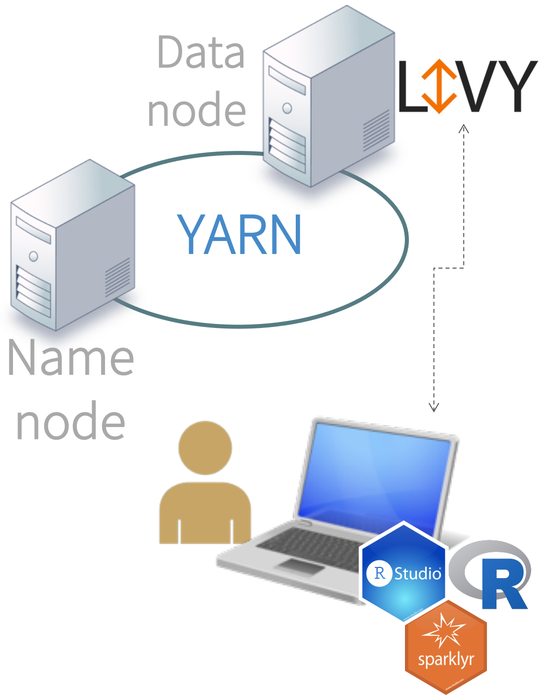
Managed Cluster



Standalone Cluster



Remote Cluster



Local

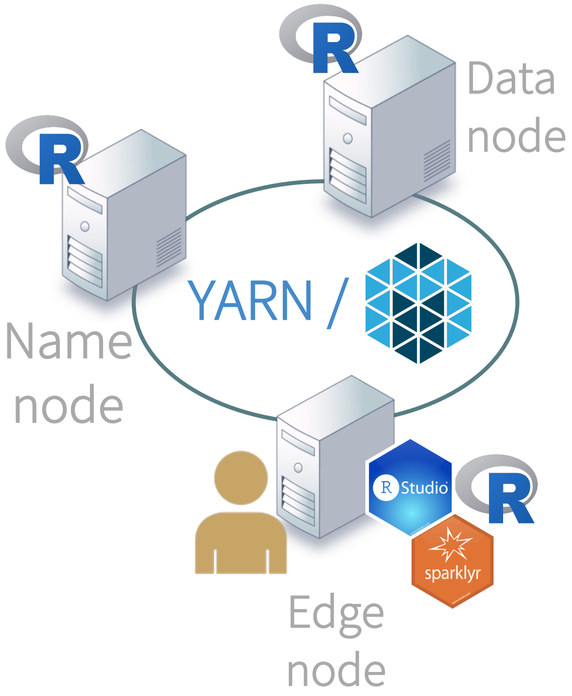
http://spark.apache.org/docs/latest/cluster-overview.html

https://spark.rstudio.com/articles/deployment-overview.html

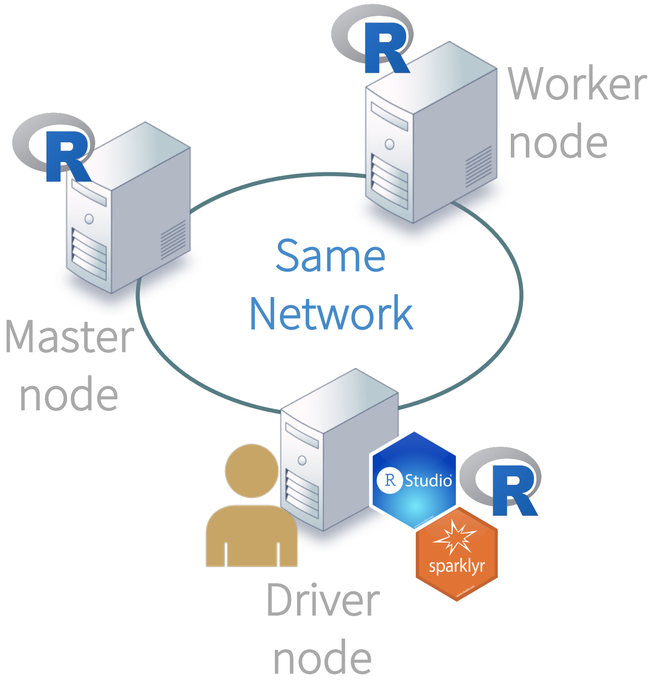
RStudio®

DEPLOYMENT MODES WITH SPARK_APPLY()

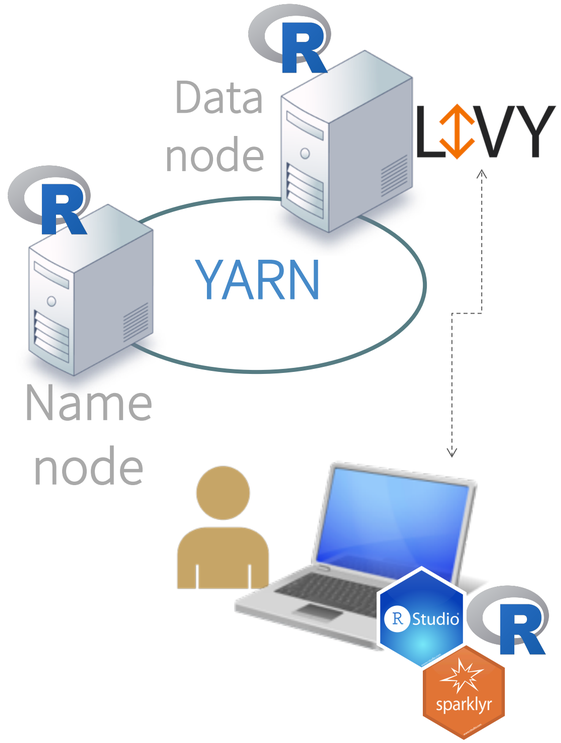
Managed Cluster



Standalone Cluster



Remote Cluster



Using spark_apply() adds the requirement to have R pre-installed in all of the nodes of the cluster

https://spark.rstudio.com/articles/guides-distributed-r.html

RStudio®

Spark Concepts

Driver

A program in a node

A Driver program starts a Spark context. One node can start multiple Driver programs

Master

Cluster manager

A node (server)

YARN

Cluster service

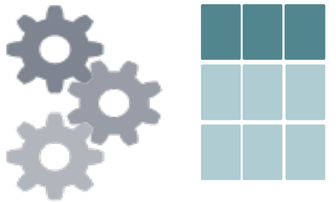
One Master can exist in the cluster

Worker

A node (server)

Multiple worker nodes can exists in the cluster

Executor



Processes that run computations and store data

They run on a Worker or Driver node. One Worker node can have multiple Executors

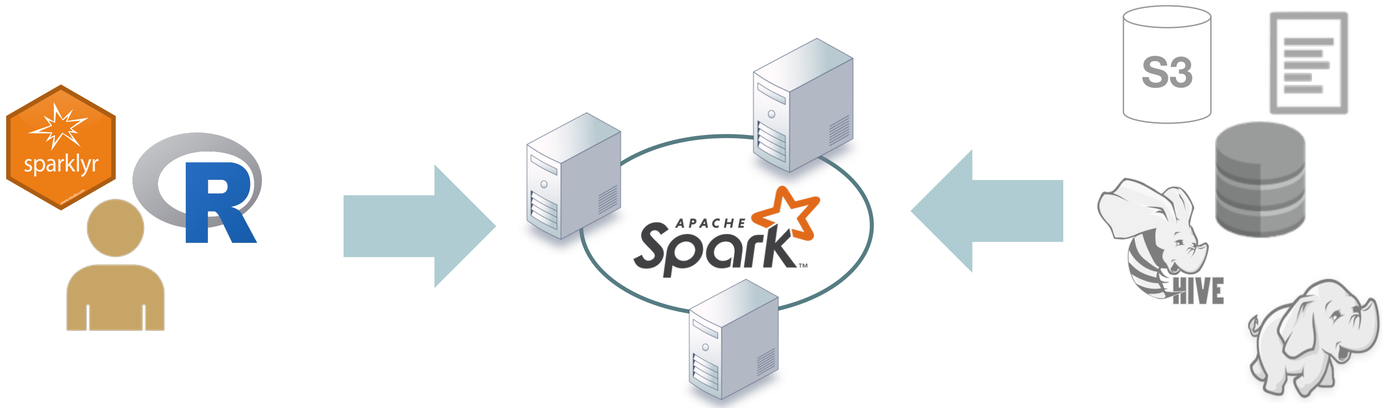
https://spark.apache.org/docs/latest/cluster-overview.html

RStudio®

DATA LOAD OPTIONS WITH SPARKLYR

Option 1

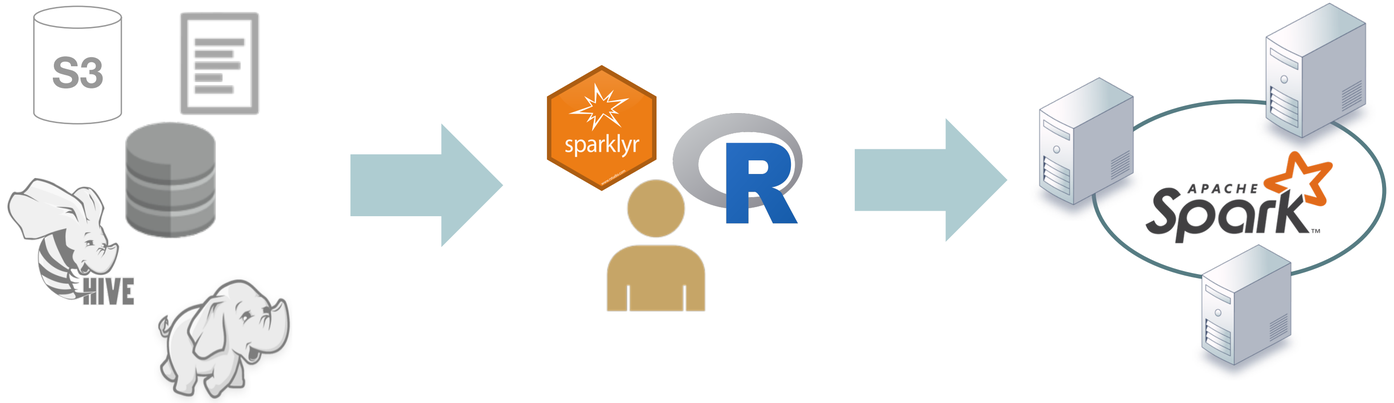
- Use sparklyr to tell the cluster where the data is located
- Spark maps and reads the data source



- Best approach for large datasets
- Depends on Spark packages. All nodes must be able to access the package using the same path (URL, network path, etc.)
- All nodes must have access to the file(s) using the same path (URL, network path, etc.)
- For Standalone mode, a simple approach is to create a network share amongst the nodes

Option 2

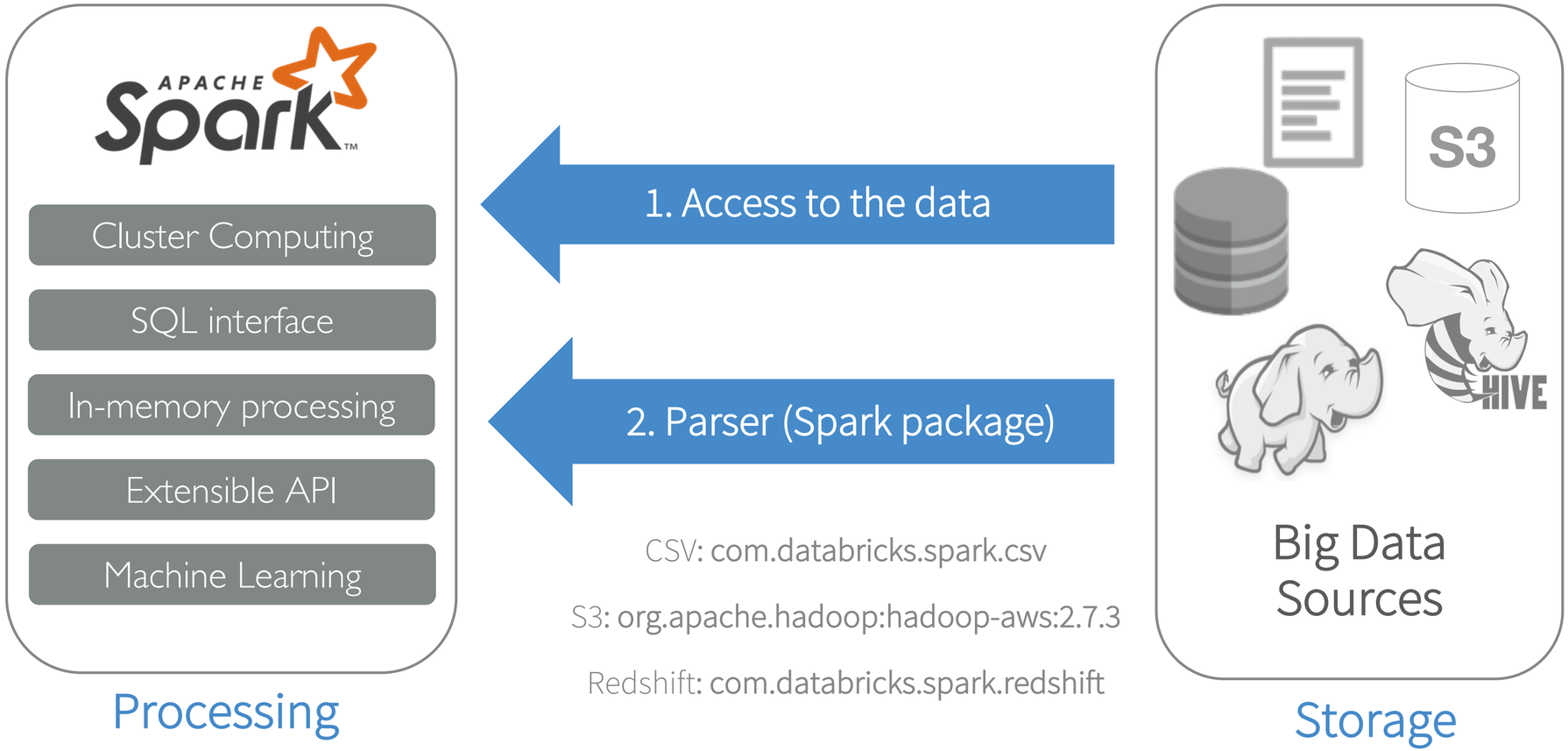
- Data is read into R
- Data is pushed to the cluster using *copy_to()*



- Great approach for small datasets
- May be the only option if all of the nodes do not have access to the file

RStudio®

2 THINGS SPARK NEEDS TO READ DATA



RStudio®

Reference Links

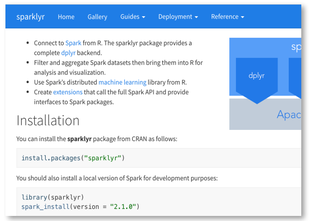
Configuring Spark Connections (Added 8/25/17)
https://spark.rstudio.com/articles/deployment-connections.html

Spark Standalone Deployment in AWS (Updated 8/25/17)
https://spark.rstudio.com/articles/deployment-amazon-ec2.html

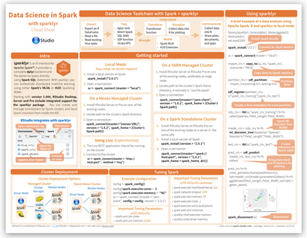
Spark Standalone Mode and S3 (Will be updated by 9/30/17)
https://spark.rstudio.com/articles/deployment-amazon-s3.html

RStudio®

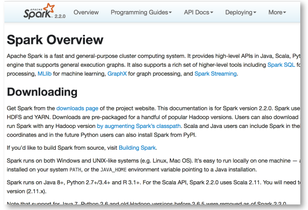
Useful Links



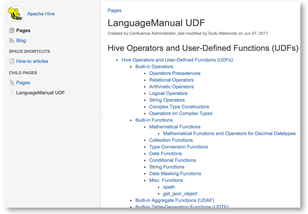
sparklyr's official website
spark.rstudio.com



sparklyr, IDE, RMarkdown cheatsheets
www.rstudio.com/resources/cheatsheets



Spark documentation
https://spark.apache.org/docs/latest/



Hive SQL UDF
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF

RStudio®

RStudio®

Open Source & Free

Desktop: http://www.rstudio.com/products/rstudio/download/
RStudio Server: http://www.rstudio.com/products/rstudio/download-server/
Shiny Server: http://www.rstudio.com/products/shiny/download-server/
shinyapps.io beta: https://www.shinyapps.io/admin/#/signup

45 Day Evaluation of Pro Products

RStudio Server Pro: http://www.rstudio.com/products/rstudio-server-pro/evaluation/
Shiny Server Pro: http://www.rstudio.com/products/shiny-server-pro/evaluation/

PLEASE STAY IN TOUCH

Blog - http://rviews.rstudio.com/

Blog - http://blog.rstudio.org/

Twitter - @rstudio #rstats http://twitter.com/rstudio/

GitHub - https://github.com/rstudio/

LinkedIn - https://linkedin.com/company/rstudio-inc

Facebook - https://www.facebook.com/pages/RStudio-inc

Google+ - https://plus.google.com/110704473211154995841/posts

RStudio®