



# sparklyr

## Understanding Spark and sparklyr deployment modes

The screenshot shows the R Studio interface with the following components:

- Top Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help.
- File Tab:** chicagoFood.R x
- Code Editor:** Contains R code for reading a CSV file from a Chicago government API and creating a Leaflet map.

```
1 url <- "http://data.cityofchicago.org/api/views/4ijn-s7e5/rows.csv?c"
2 data <- read.csv(url, header = TRUE) # takes a minute...
3 names(data) <- tolower(names(data))
4 data1 <- subset(data, risk %in% c("Risk 1 (High)", "Risk 2 (Medium)", "Risk 3 (Low)"))
5 data1$risk <- droplevels(data1$risk)
6
7 data1 <- data1[1:50,]
8 library(leaflet)
9 leaflet(data1) %>%
10   addTiles() %>%
11   addMarkers(lat = ~latitude, lng = ~longitude)
```
- Console:** Shows the R code being run and its output.

```
> data1 <- subset(data, risk %in% c("Risk 1 (High)", "Risk 2 (Medium)", "Risk 3 (Low)"))
> data1$risk <- droplevels(data1$risk)
>
> data1 <- data1[1:50,]
> library(leaflet)
> leaflet(data1) %>%
+   addTiles() %>%
+   addMarkers(lat = ~latitude, lng = ~longitude)
```
- Environment:** Shows the global environment with two datasets: `data` (118607 obs. of 17 variables) and `data1` (50 obs. of 17 variables).
- Plots:** A Leaflet map of Chicago showing markers at various locations across the city.

# SPARKLYR - AUTHORS



Javier Luraschi – author , creator



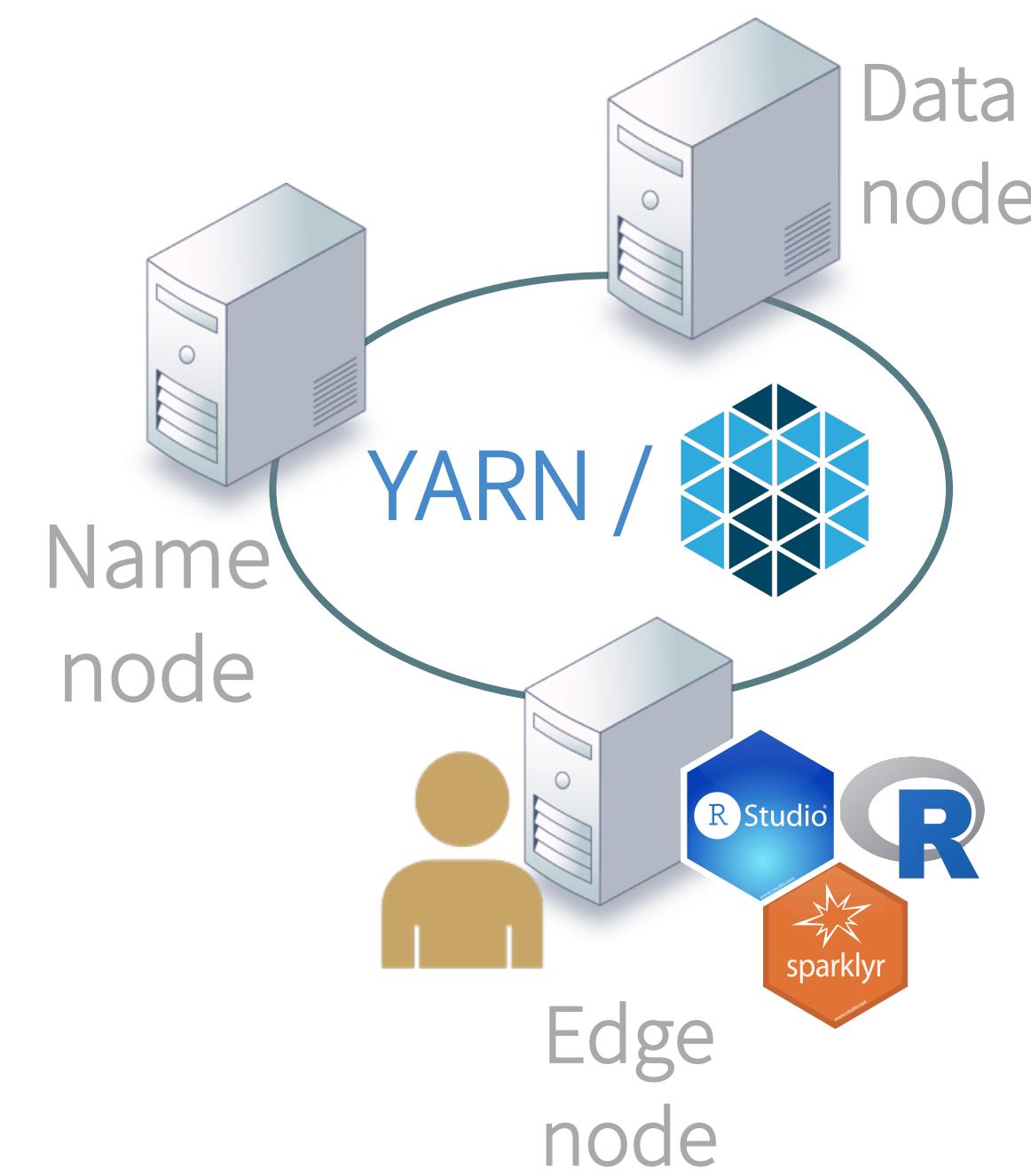
JJ Allaire – author



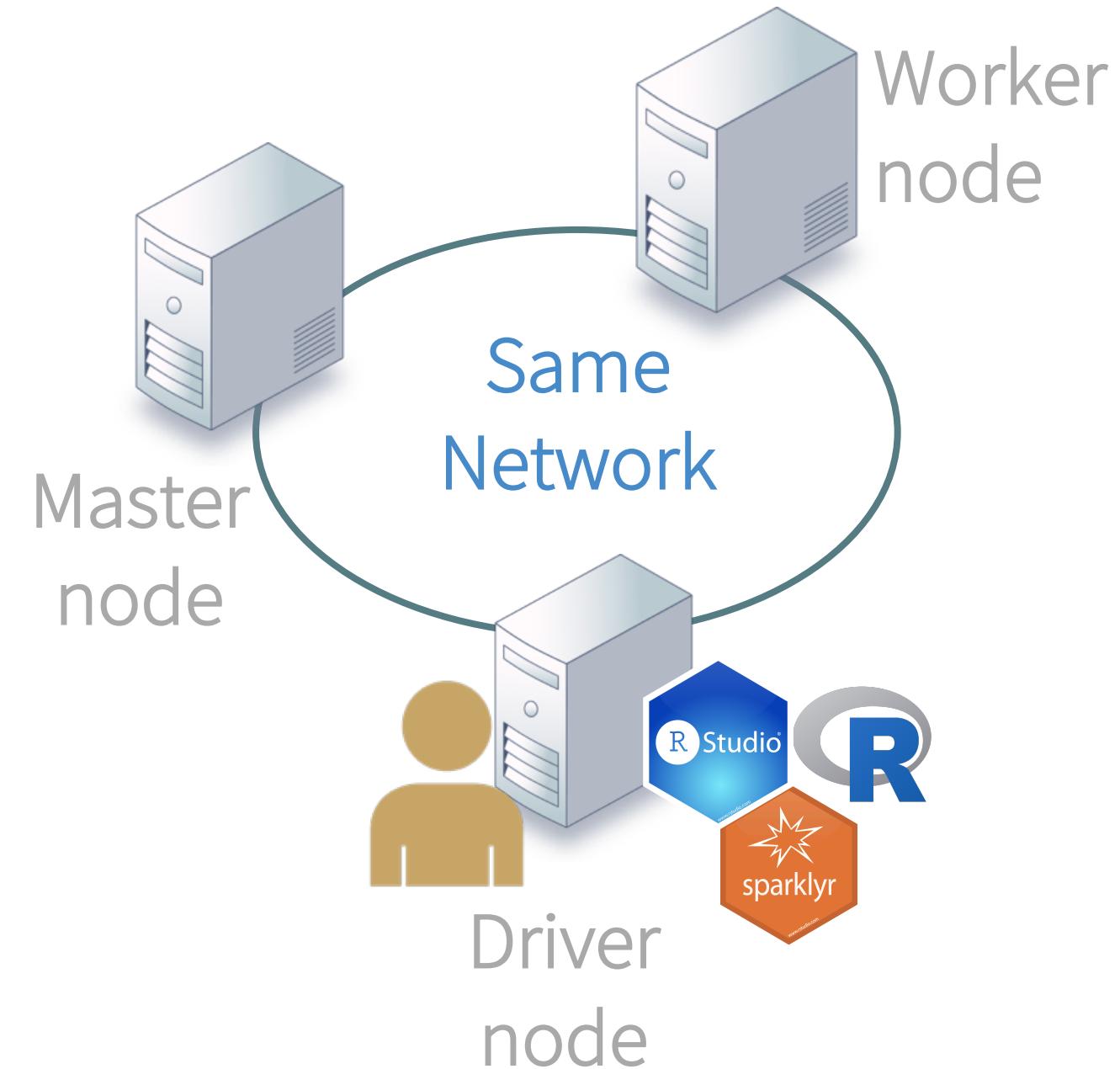
Kevin Ushey – author

# DEPLOYMENT MODES

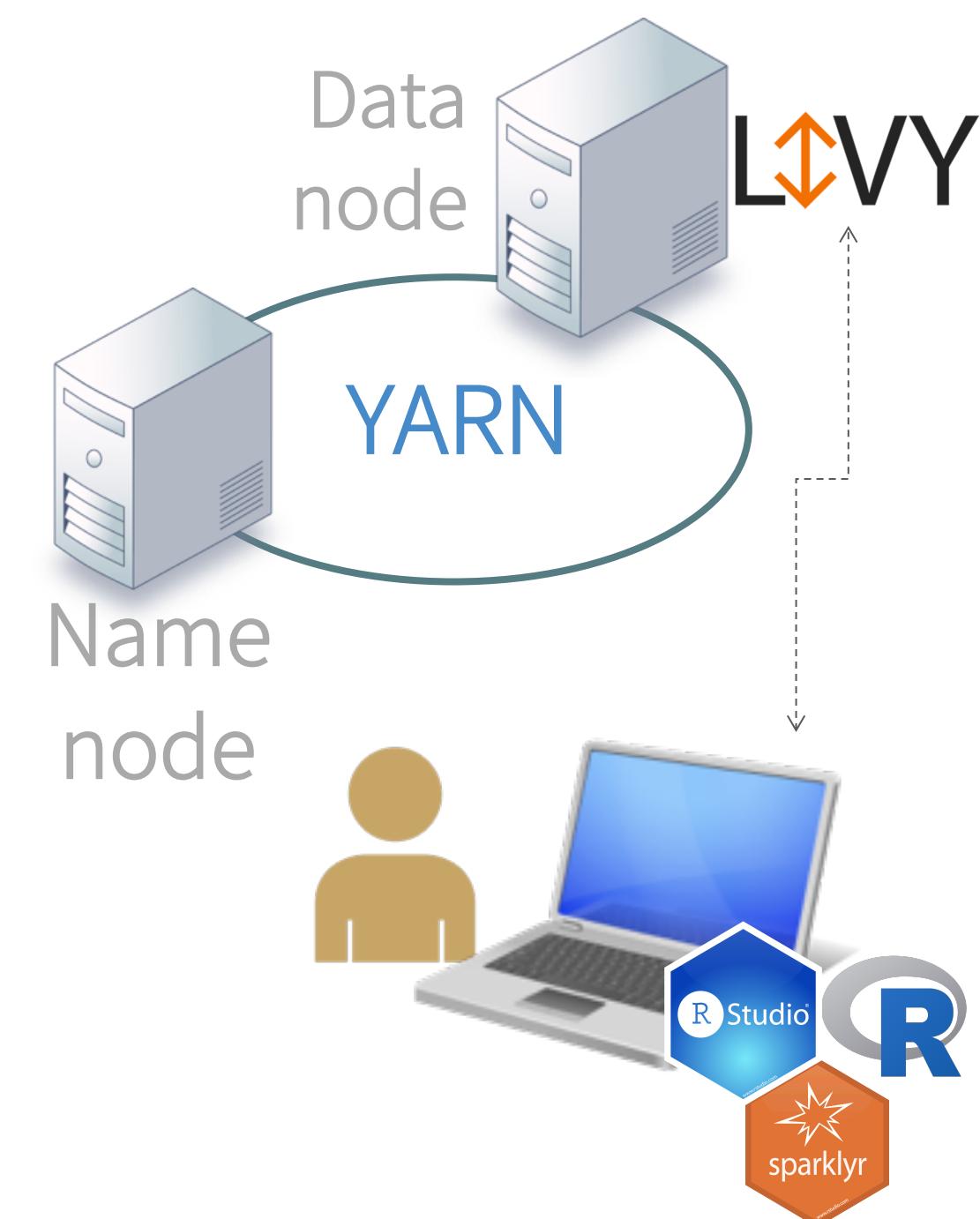
Managed Cluster



Standalone Cluster



Remote Cluster



Local

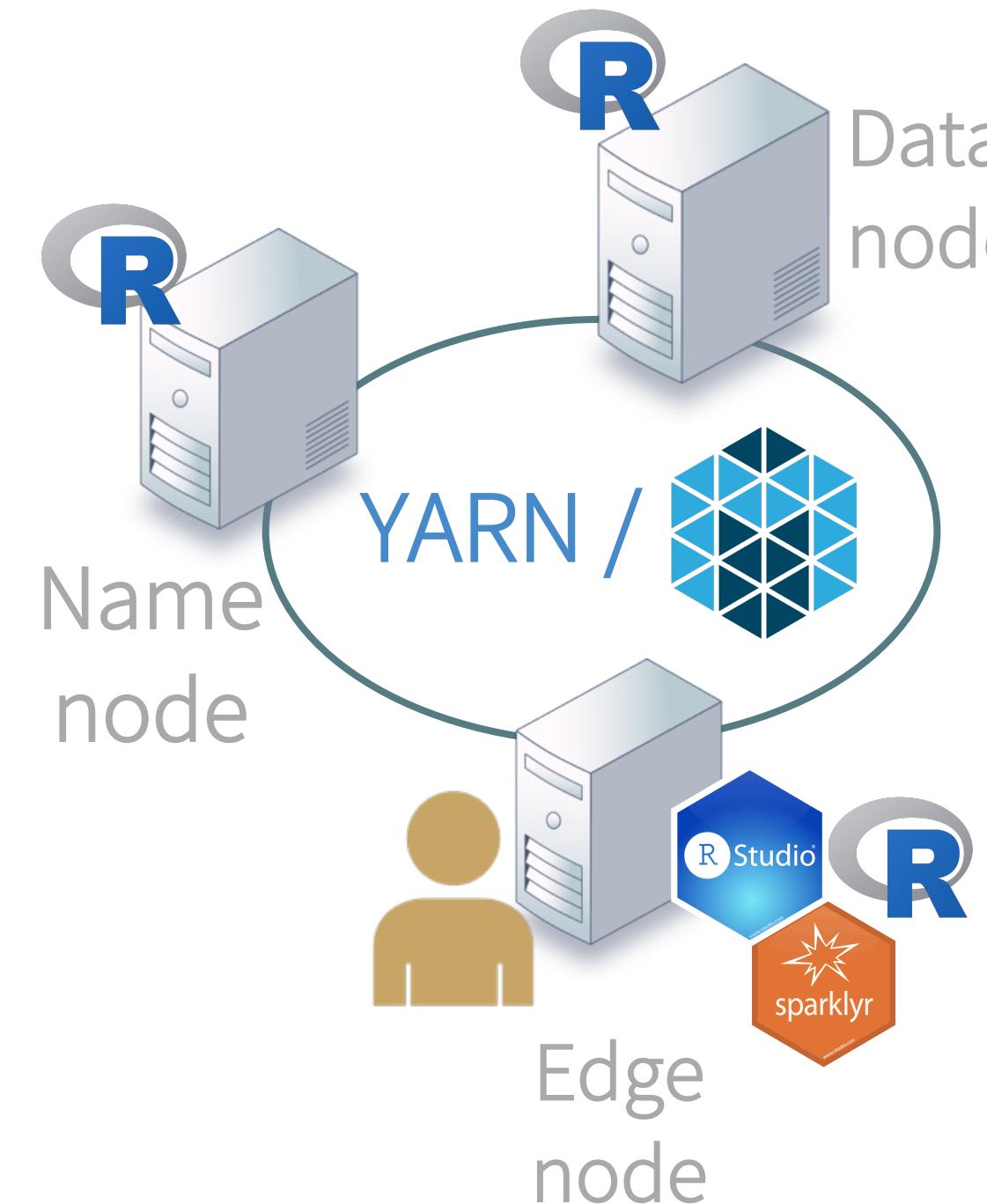


<http://spark.apache.org/docs/latest/cluster-overview.html>

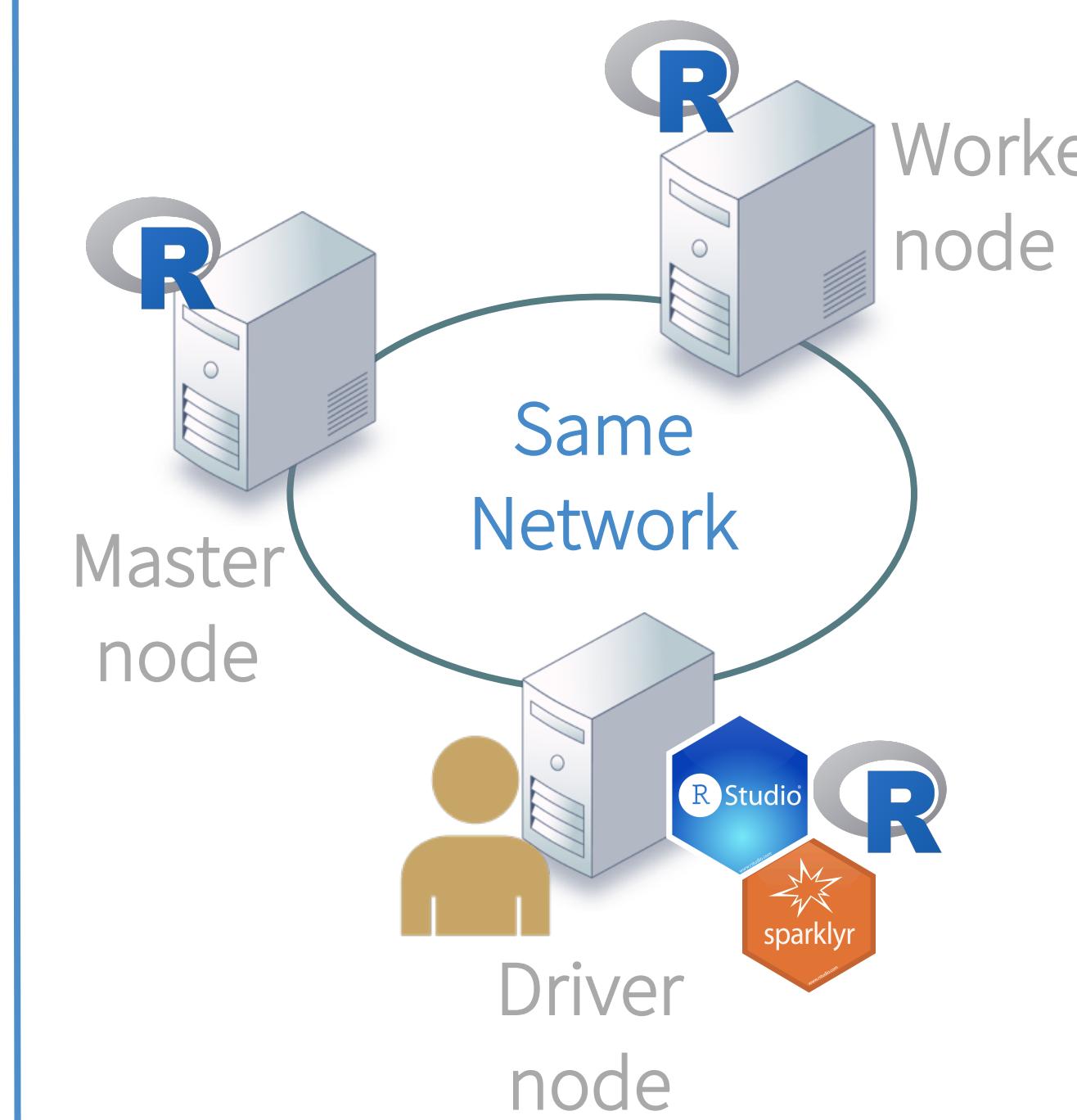
<https://spark.rstudio.com/articles/deployment-overview.html>

# DEPLOYMENT MODES WITH SPARK\_APPLY()

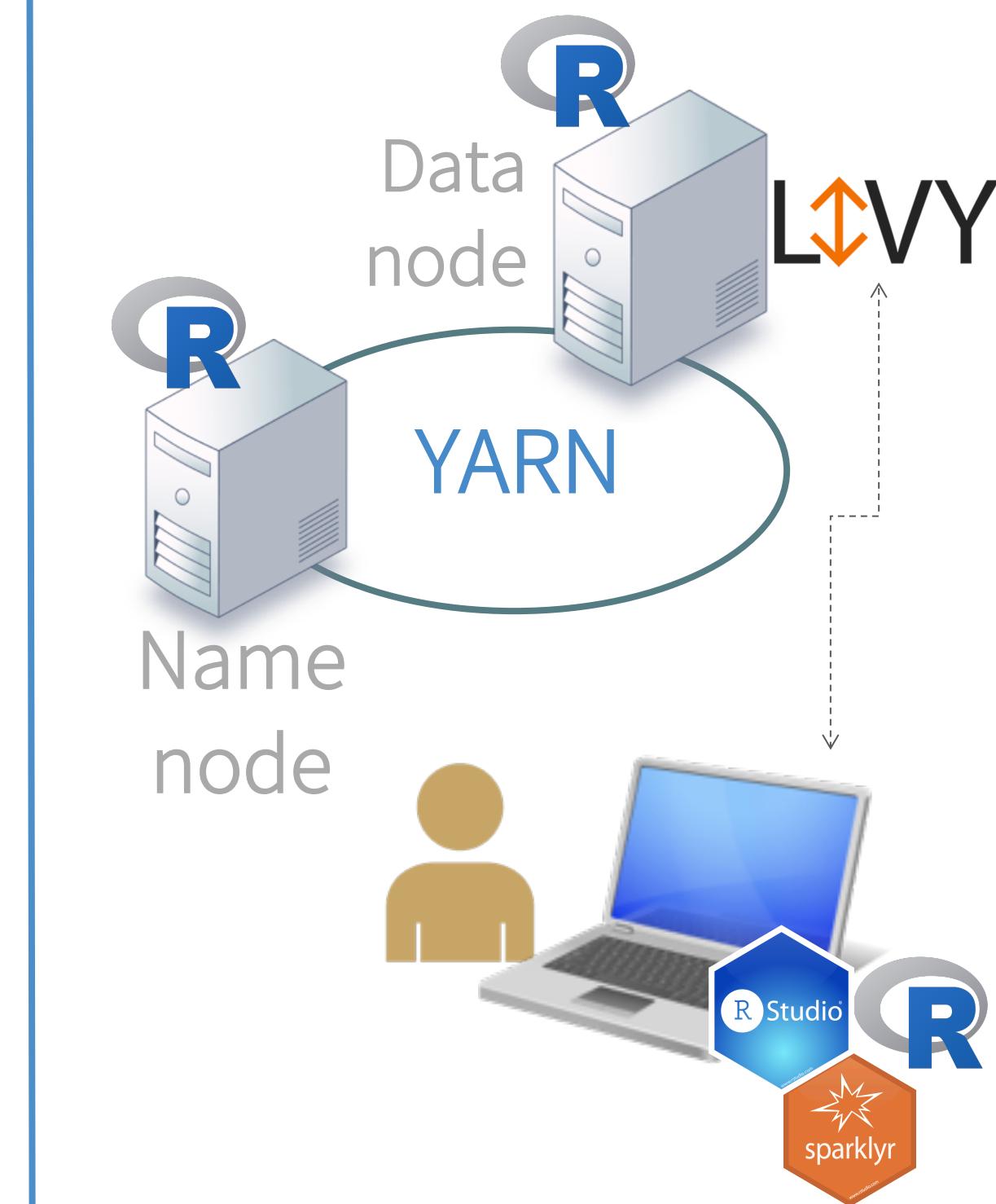
Managed Cluster



Standalone Cluster



Remote Cluster

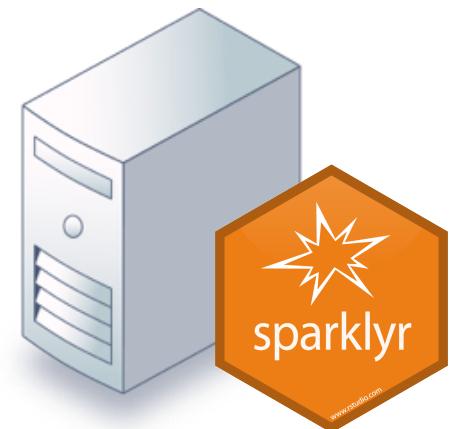


Using `spark_apply()` adds the requirement to have R pre-installed in all of the nodes of the cluster

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

# SPARK CONCEPTS

## Driver

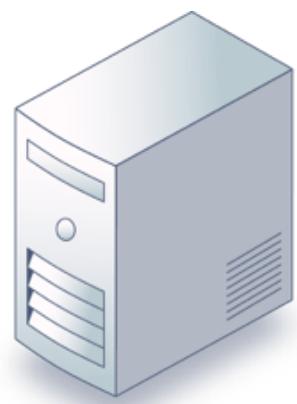


A program in a node

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

## Master

Cluster manager



YARN

A node (server)

Cluster service

One Master can exist in the cluster

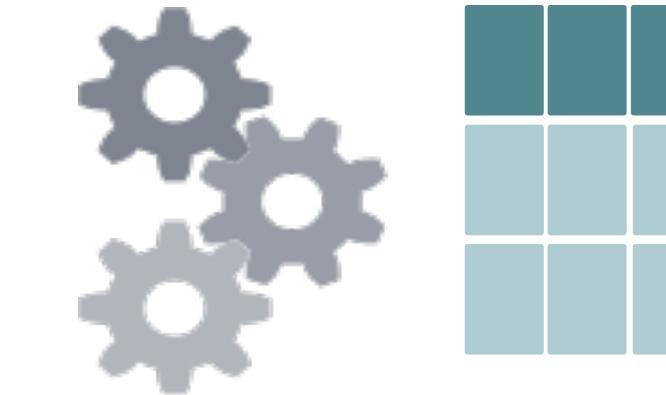
## Worker



A node (server)

Multiple worker nodes can exist in the cluster

## Executor



Processes that run computations and store data

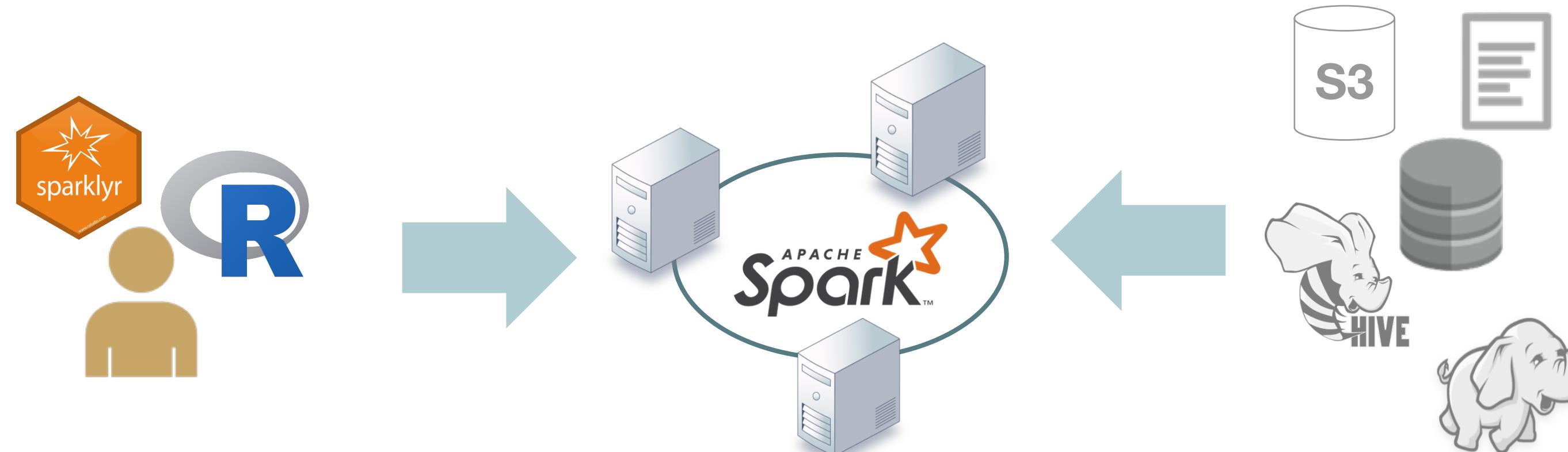
They run on a Worker or Driver node. One Worker node can have multiple Executors

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

# 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

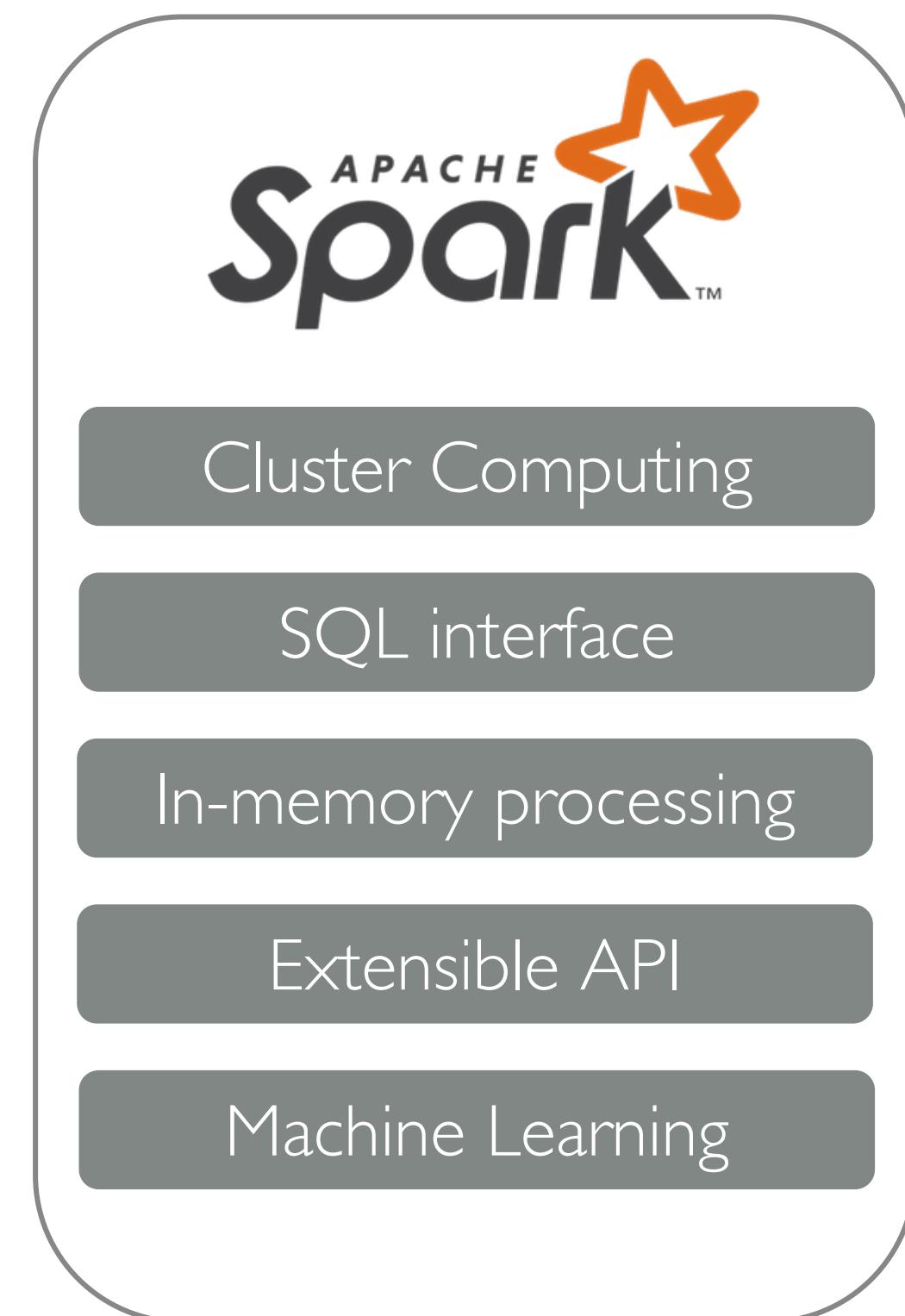
## Option 2

- 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

# 2 THINGS SPARK NEEDS TO READ DATA



Processing

1. Access to the data
2. Parser (Spark package)

CSV: com.databricks.spark.csv  
S3: org.apache.hadoop:hadoop-aws:2.7.3  
Redshift: com.databricks.spark.redshift



Storage

# REFERENCE LINKS

**Configuring Spark Connections** (Added 8/25/17)

<https://spark.rstudio.com/articles/deployment-connections.html>

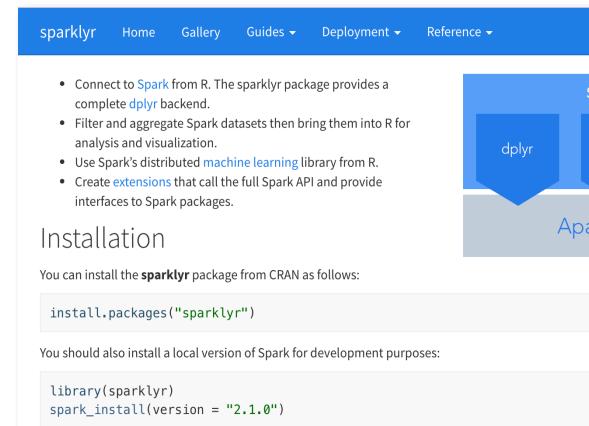
**Spark Standalone Deployment in AWS** (Updated 8/25/17)

<https://spark.rstudio.com/articles/deployment-amazon-ec2.html>

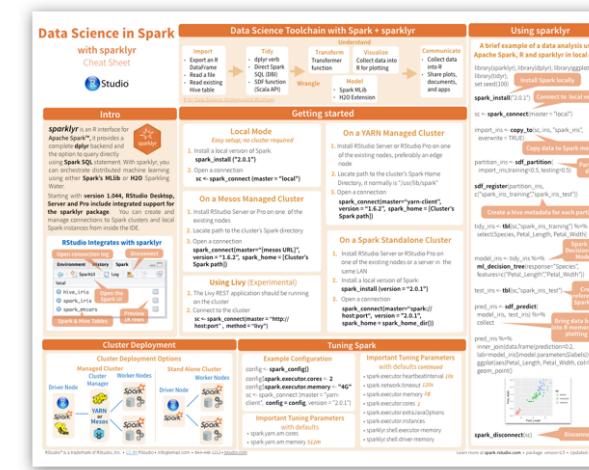
**Spark Standalone Mode and S3** (Will be updated by 9/30/17)

<https://spark.rstudio.com/articles/deployment-amazon-s3.html>

# USEFUL LINKS



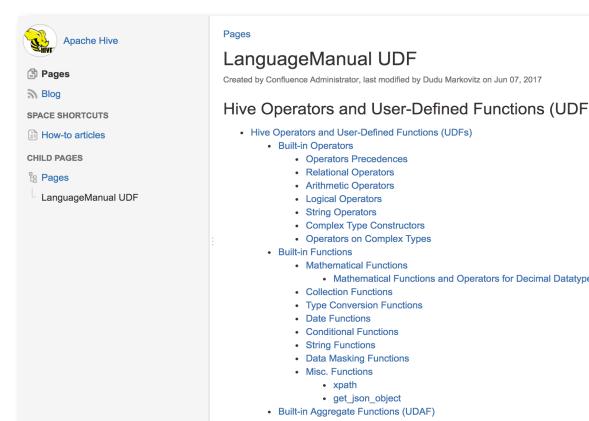
sparklyr's official website  
[spark.rstudio.com](http://spark.rstudio.com)



sparklyr, IDE, RMarkdown cheatsheets  
[www.rstudio.com/resources/cheatsheets](http://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>



## **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>