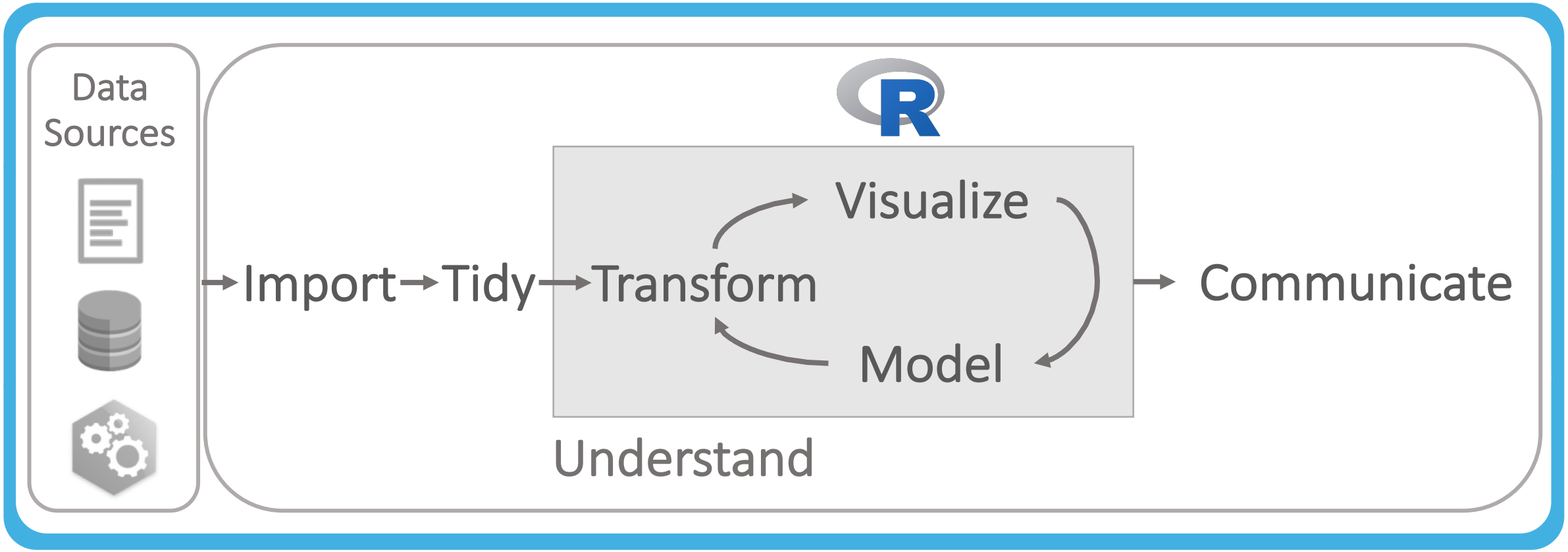
sparklyr – *An R interface for Apache Spark*

Edgar Ruiz – March 2017

R for Data Science



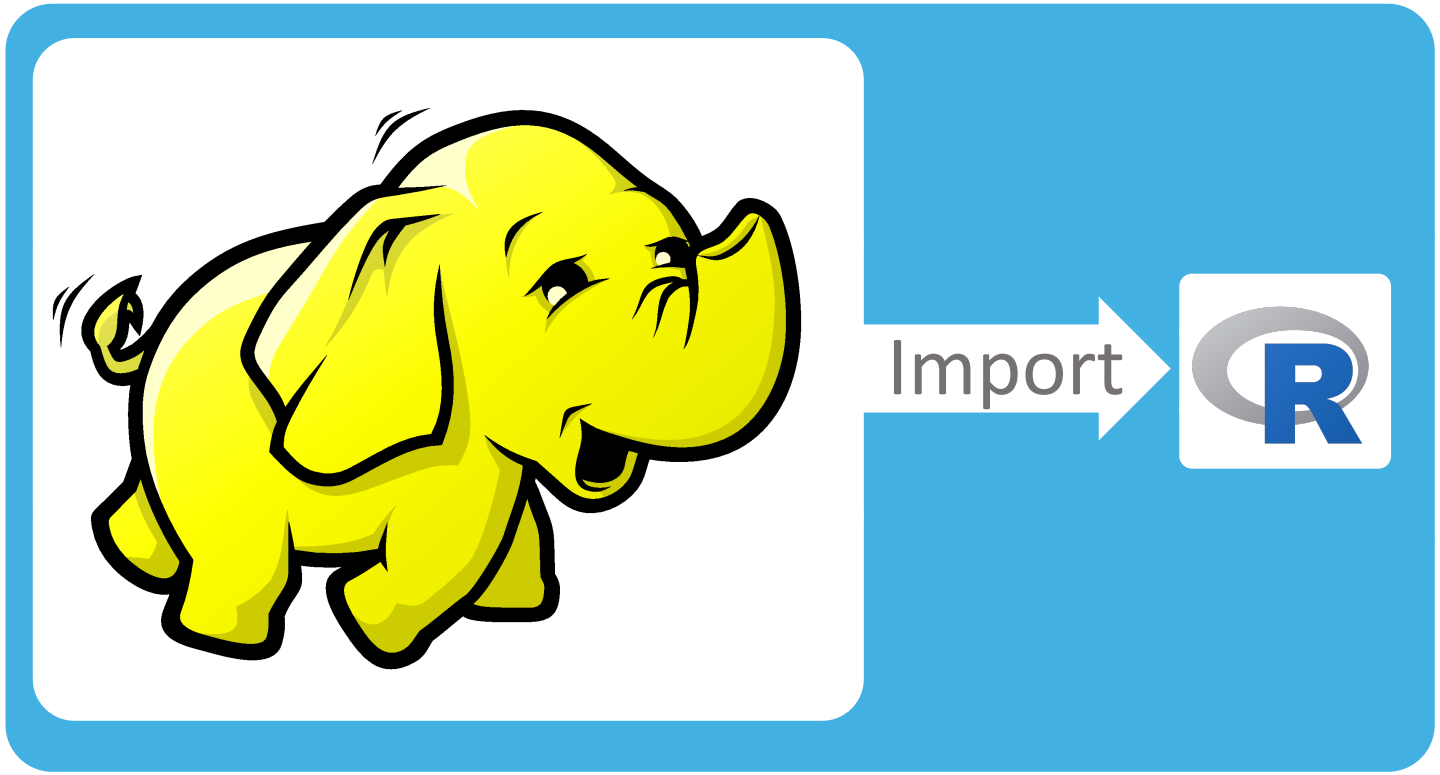
R for Data Science, Wickham & Grolemund

Use Case

Data Science using a Data Lake

Hadoop as a Data Source

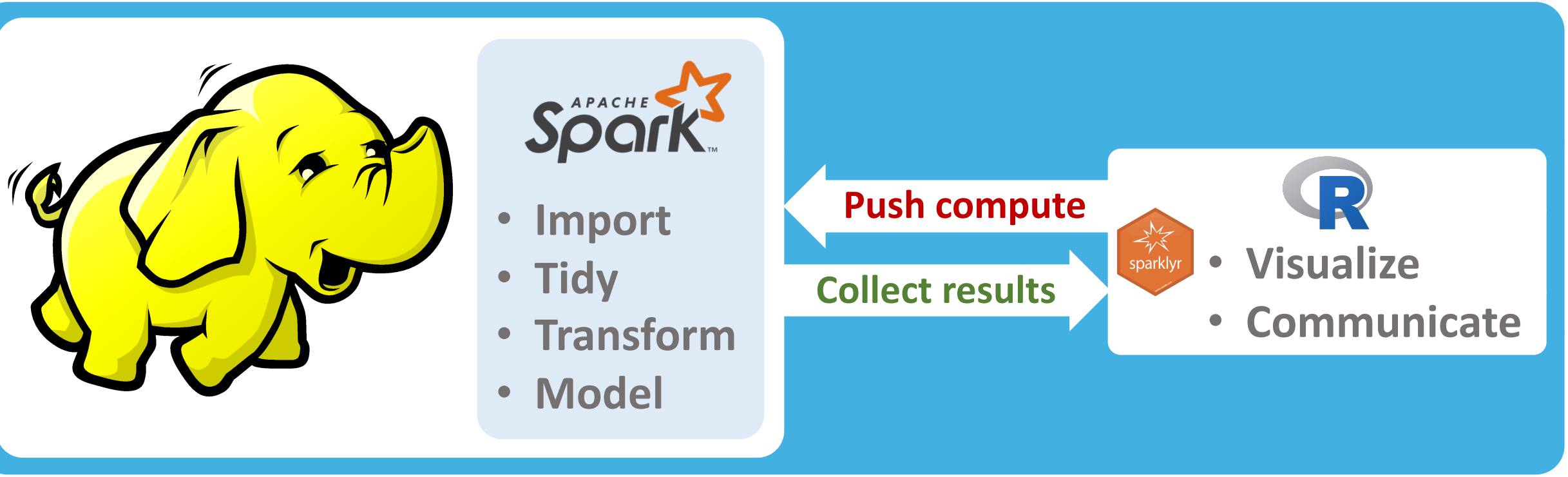
Problem: Data is too large to download into memory



Workaround: Use a very small sample or download as much data as possible

Spark as an Analysis Engine

Solution: Use sparklyr to access & analyze the data inside Spark.
Only bring results into R.

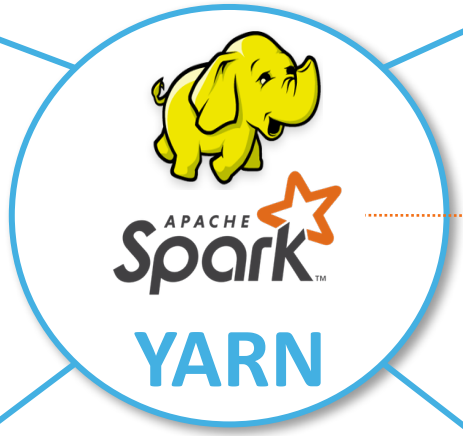


Cluster setup

Access RStudio using a web browser

Name node

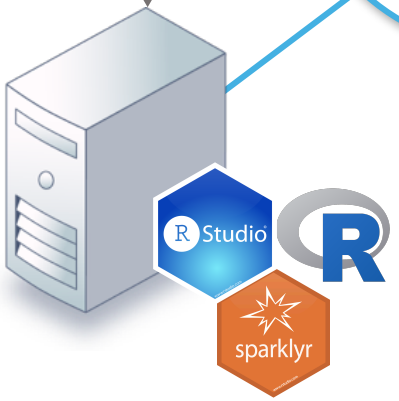
Data node



All nodes need YARN & Spark Gateway services

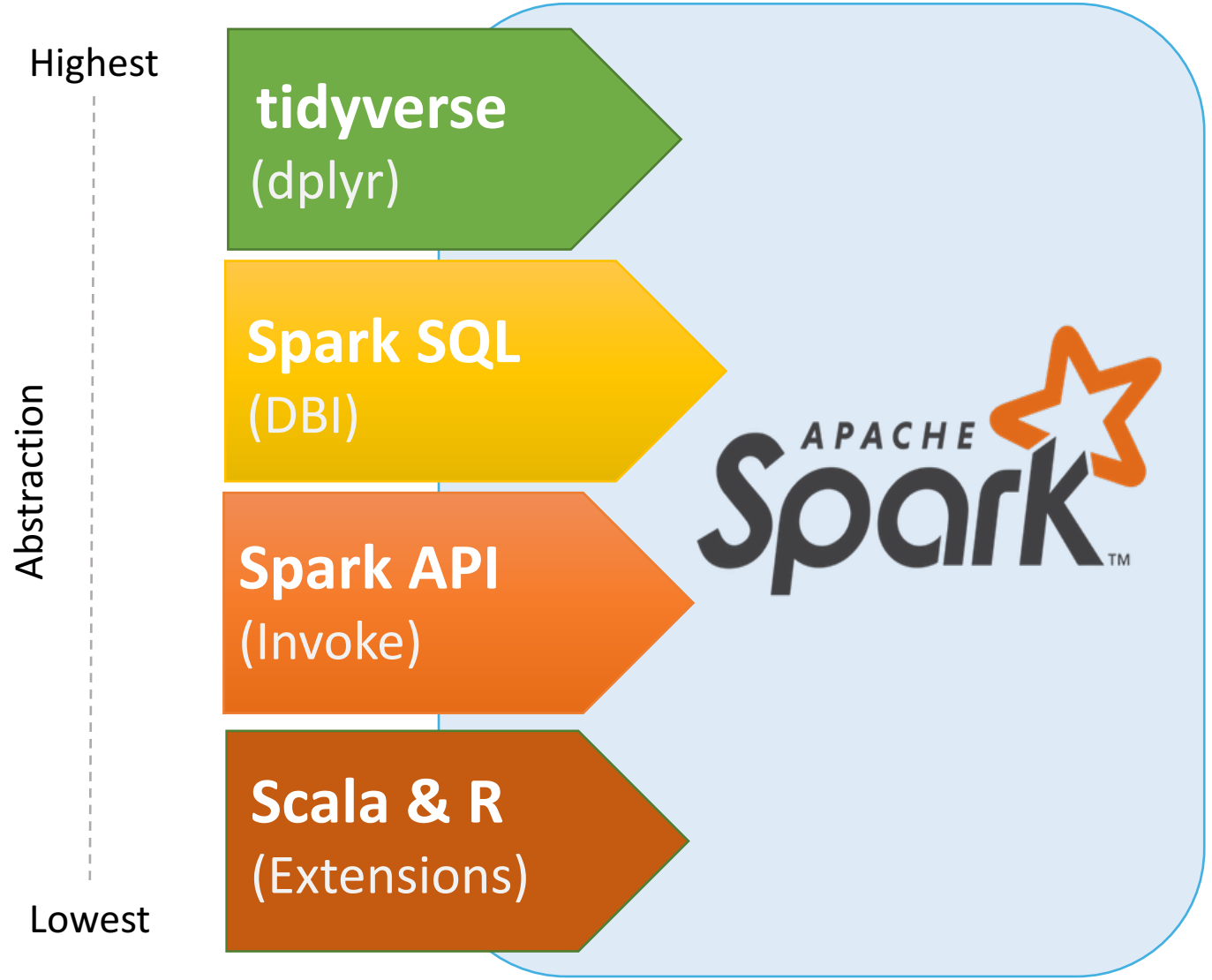
R, RStudio & sparklyr in 1 node

Edge node



Data node

4 ways to access Spark using sparklyr



Why sparklyr?

1. R, RStudio & sparklyr are needed in 1 node only

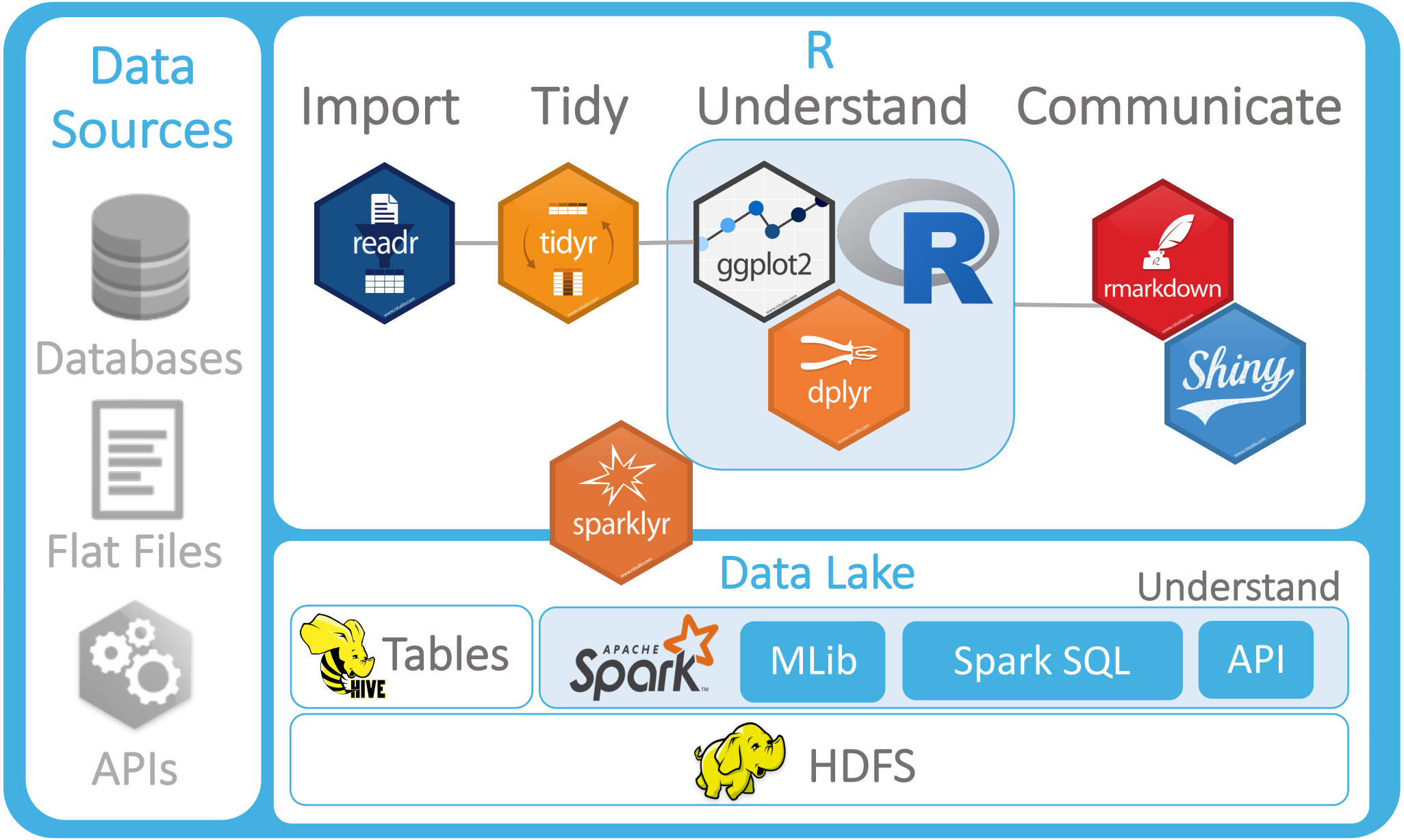
2. Access the Spark's ML library

3. Access the Spark's API framework

4. Access Hive tables & Hive's UDF

5. Interact with Spark using familiar dplyr commands

Using Spark & R for Data Science

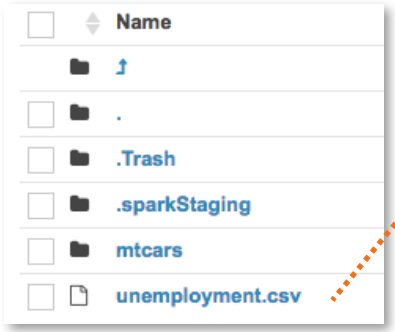


Demo

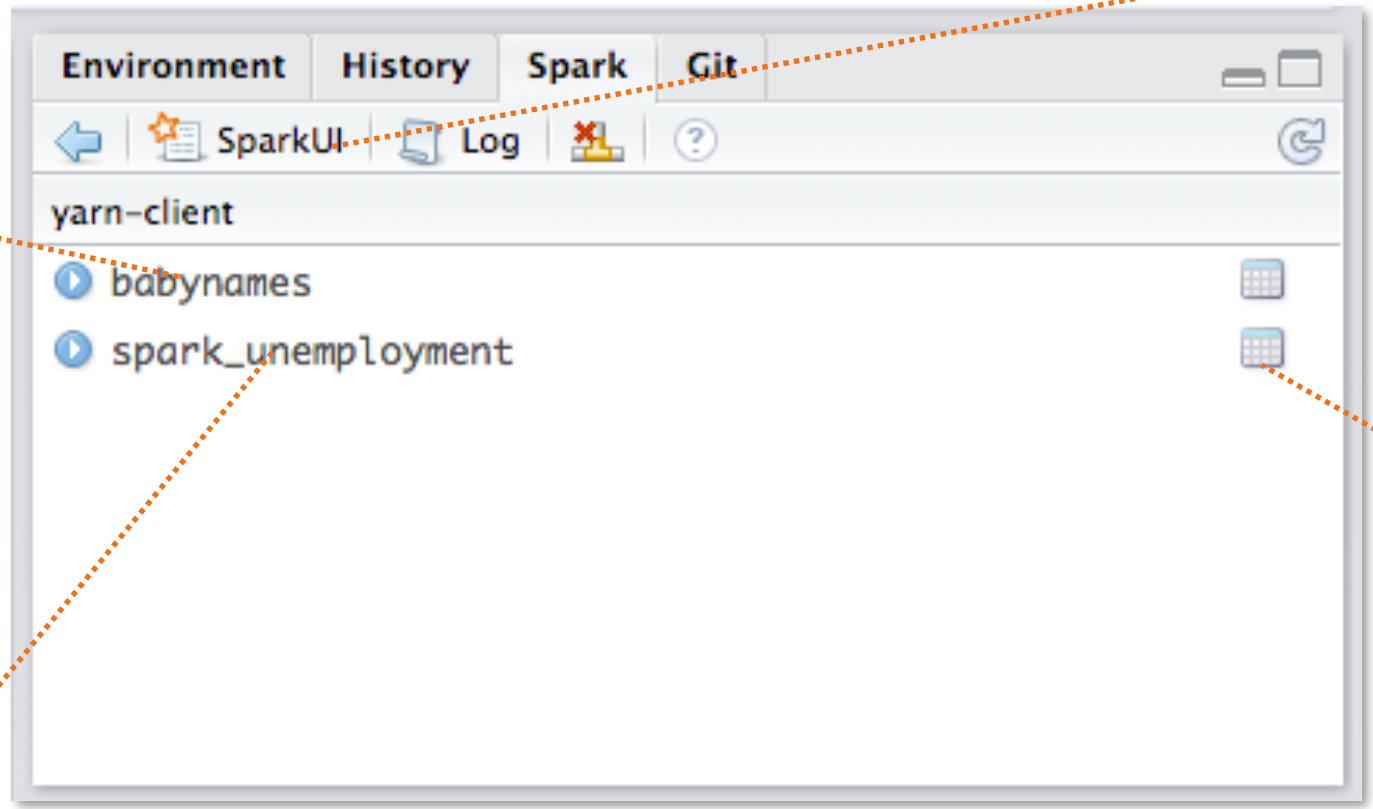
More about sparklyr

sparklyr and RStudio IDE integration

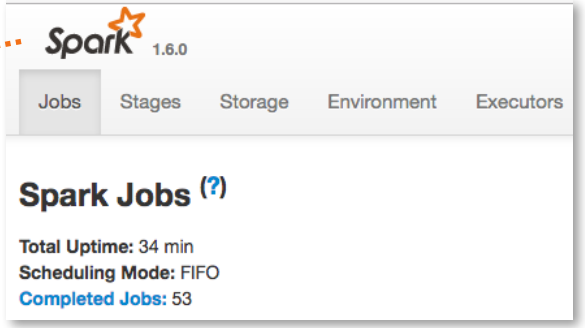
View Hive tables in R



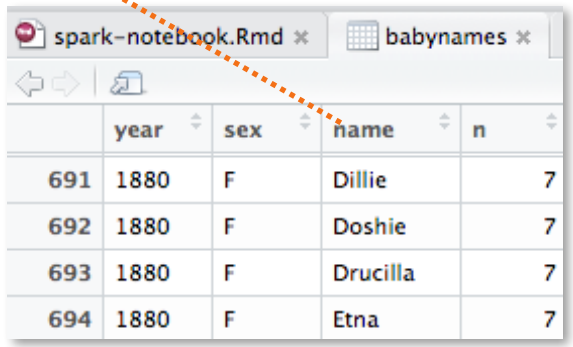
See tables in Spark session



Spark Tab inside the **RStudio IDE**



Button access to Spark UI

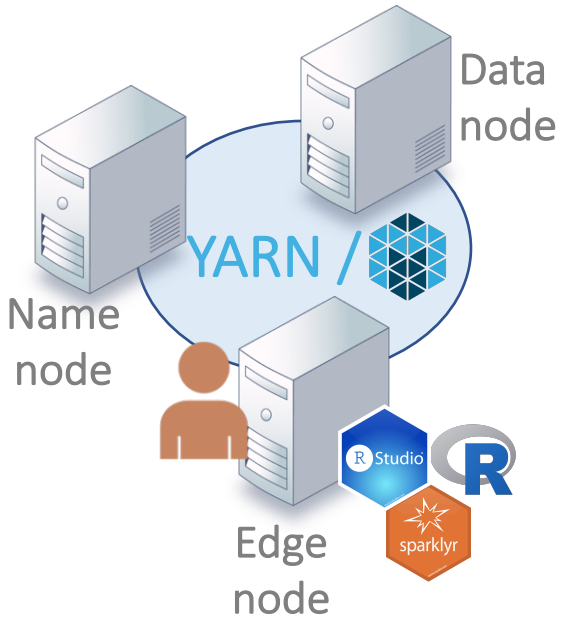


	year	sex	name	n
691	1880	F	Dillie	7
692	1880	F	Doshie	7
693	1880	F	Drucilla	7
694	1880	F	Etna	7

Preview first 1K records in R

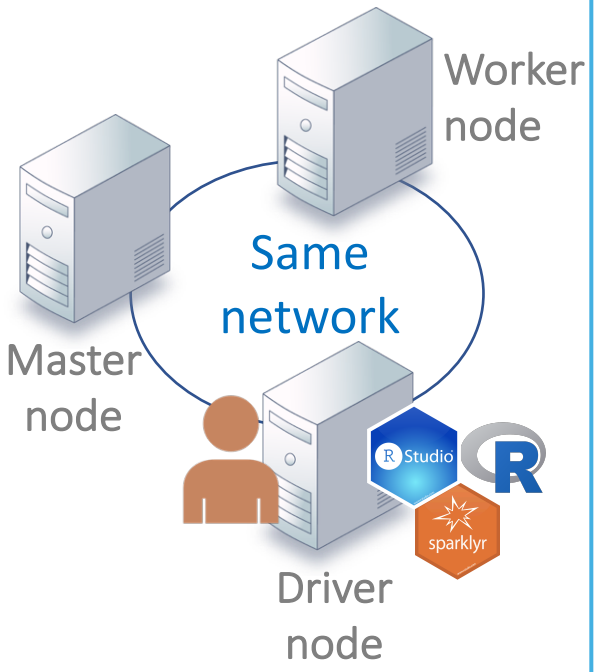
Deployment options

Managed Cluster



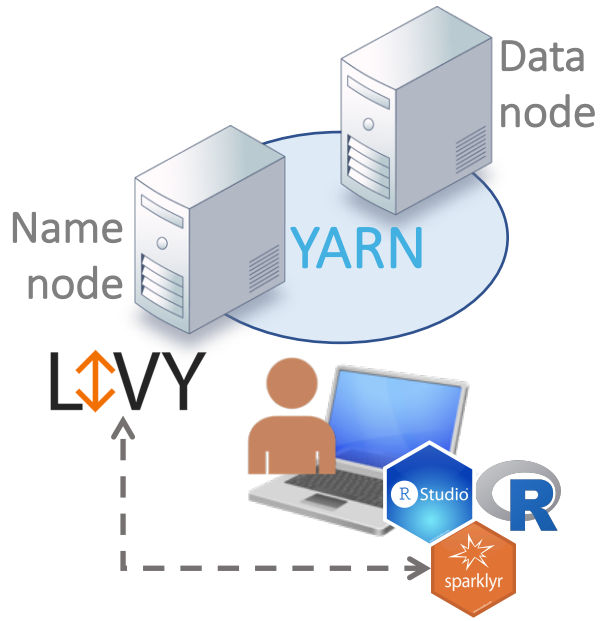
- Deployment seen at most business
- Spark version(s) available are limited to what's on the cluster

Stand Alone Cluster



- Since there's no central data repository, all data has to be either imported or connected to a common shared location (NAS, S3)

Remote Cluster with Livy

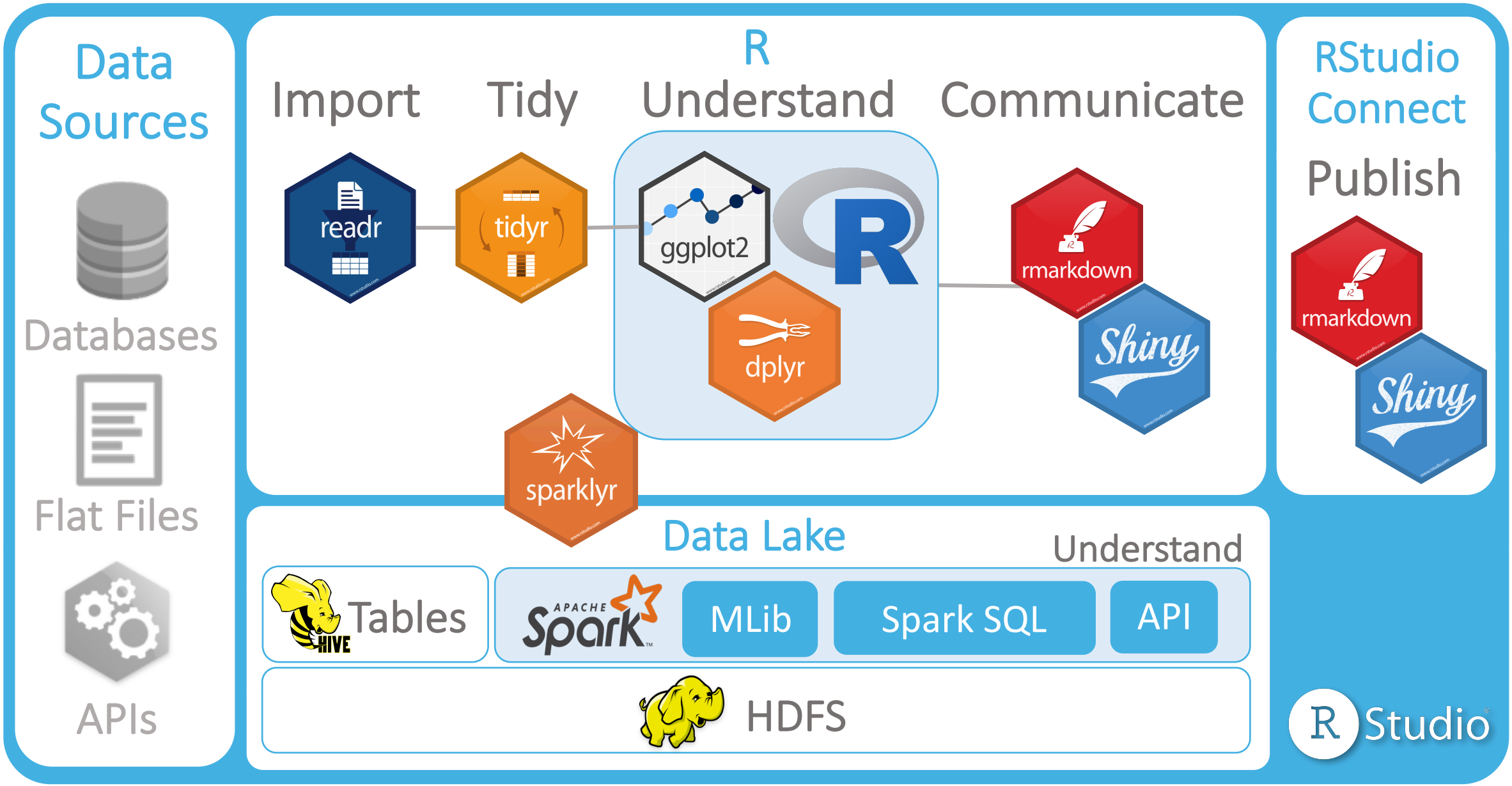


- Great for accessing a remote cluster
- New, experimental

Local

- Great for learning
- Works on Windows and Mac too
- Quick and easy way to access multiple cores

R for Data Science Toolchain with Spark



Helpful resources

Official Website	spark.rstudio.com
Data Science	http://r4ds.had.co.nz/
GitHub Repository	github.com/rstudio/sparklyr
Cheatsheet	spark.rstudio.com/images/sparklyr-cheatsheet.pdf