

Big Data with R



James
Blair

Solutions Engineer @ RStudio

Overview

1. Big Data
2. dplyr
3. dplyr + friends
4. Best practices
5. Resources



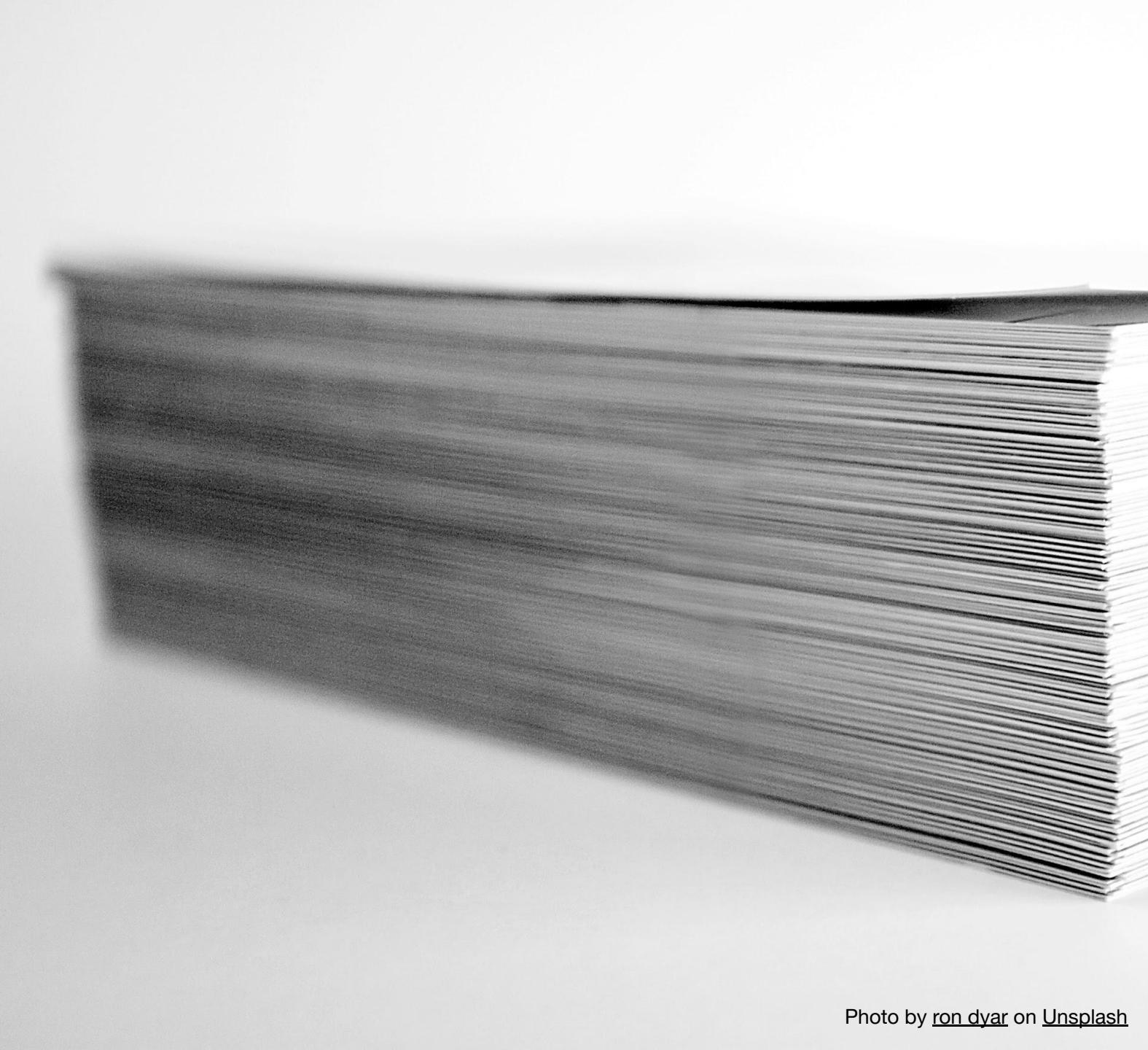
Photo by [Mat Reding](#) on [Unsplash](#)



Data > RAM

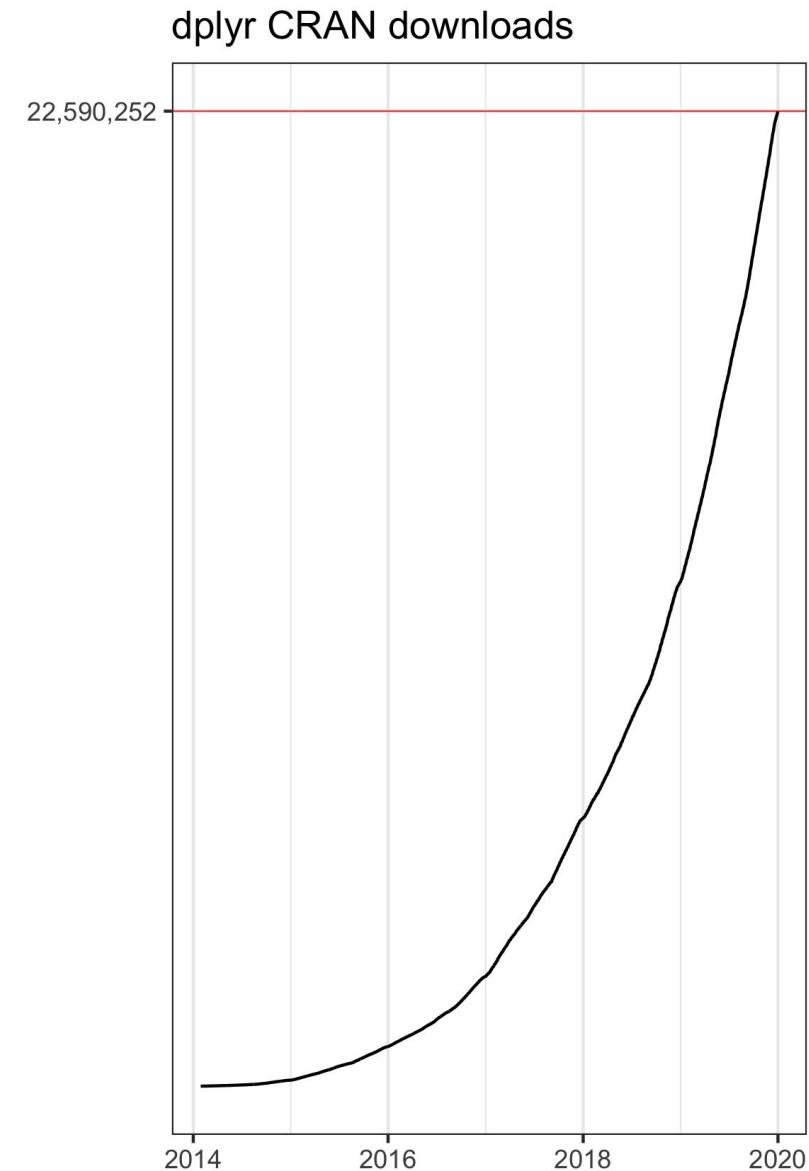
Big data forms

1. Flat file
 - a. csv
 - b. parquet
 - c. tsv
 - d. json
2. Database
3. Datalake



dplyr package

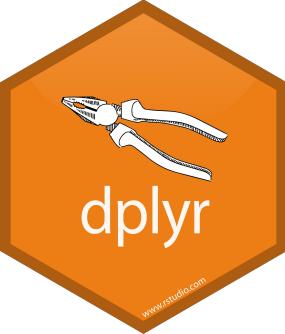
1. A grammar of data manipulation
2. Designed to **abstract over how the data is stored**
3. Consistent function interface



Data collected using the `cranlogs` R package



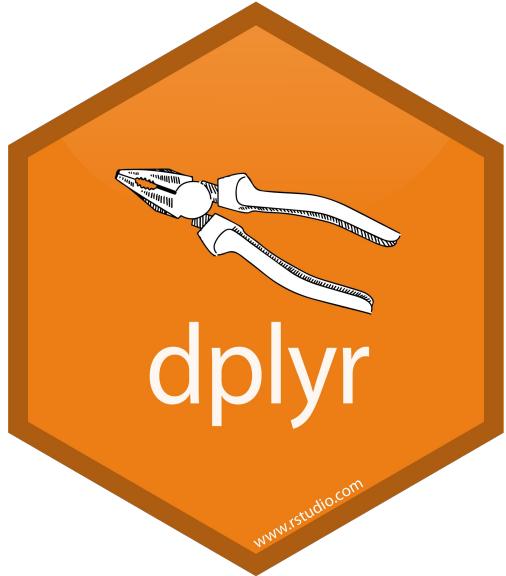
dplyr package

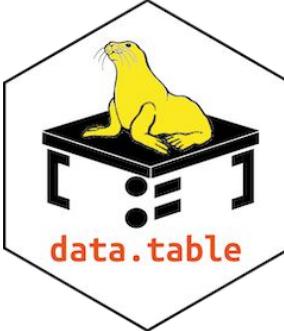


Dplyr “abstracts away how your data is stored, so that you can work with data frames, data tables and remote databases using the same set of functions.

This lets you focus on what you want to achieve, not on the logistics of data storage.

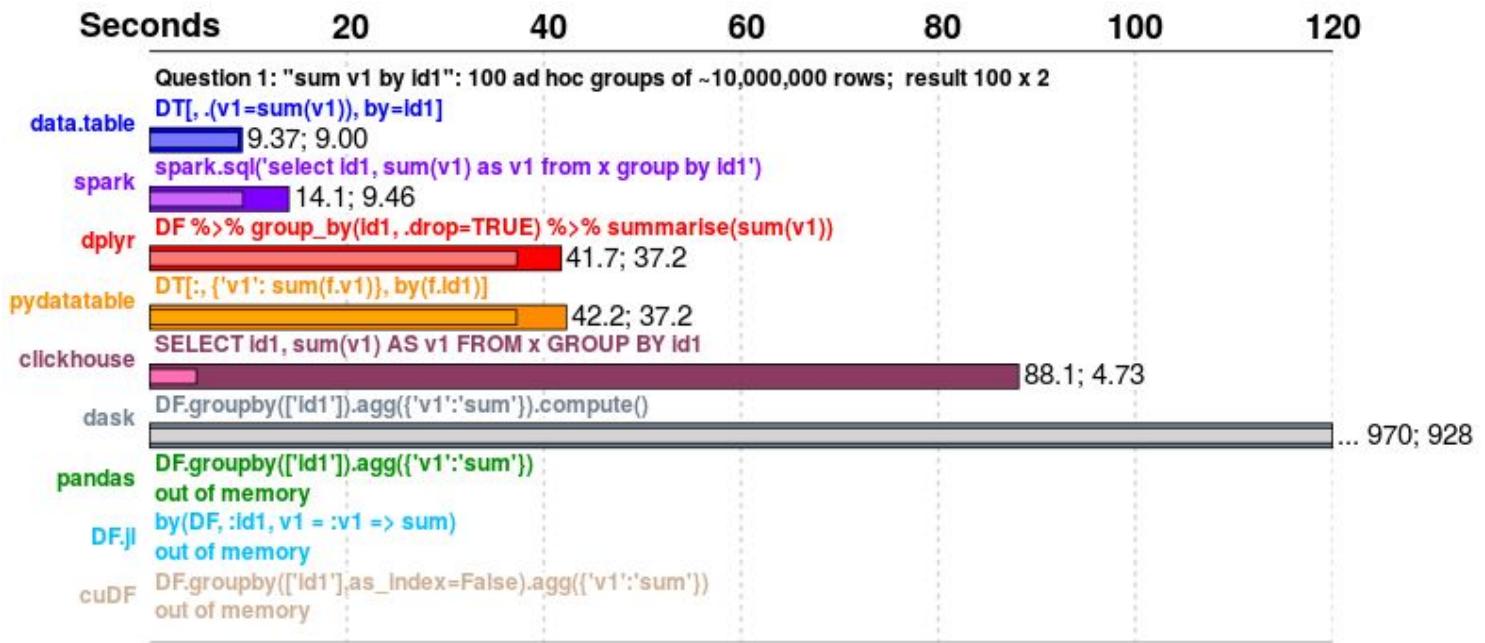
dplyr backends



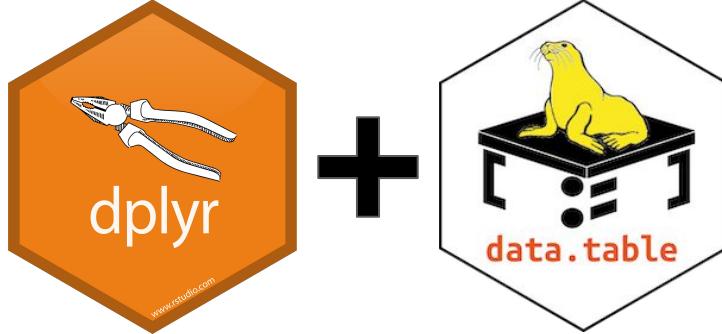


data.table package

1. High performance version of base R data.frame
2. Fast file reader fread
3. Concise syntax DT[i, j, by]



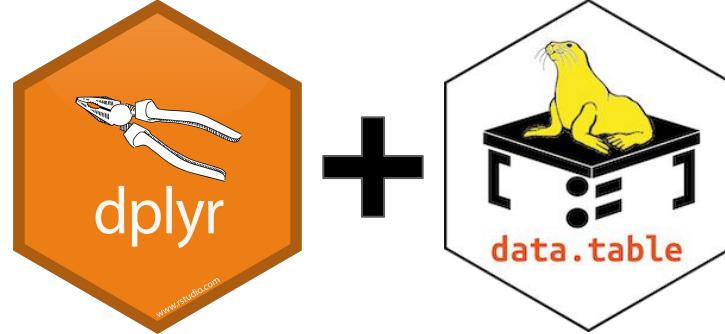
dtplyr package



The goal of dtplyr is to allow you to write dplyr code that is automatically translated to the equivalent, but usually much faster, data.table code.

dplyr package

1. Provides a `data.table` backend for `dplyr`
2. Combine the syntax of `dplyr` with the speed of `data.table`
3. Lazy evaluation
4. Converts `dplyr` syntax to `data.table` syntax

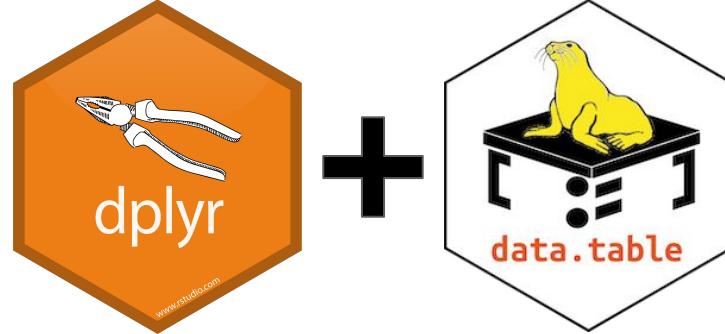


dplyr package

A word about copying...

In data.table parlance, all set functions change their input by reference. That is, no copy is made at all, other than temporary working memory, which is as large as one column.*

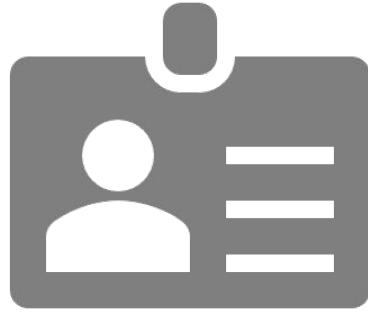
Use `lazy_dt(x, immutable = FALSE)` to prevent dplyr from making copies.



Databases



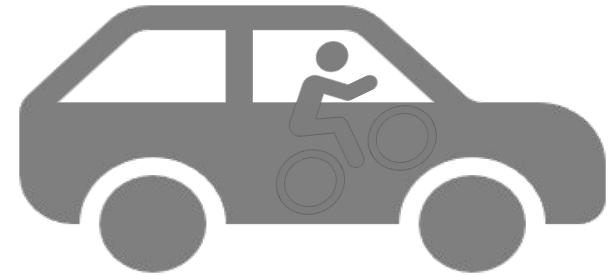
Connection requirements



Credentials



Location



Driver

Requirement definitions



- User name & password
 - Token
-

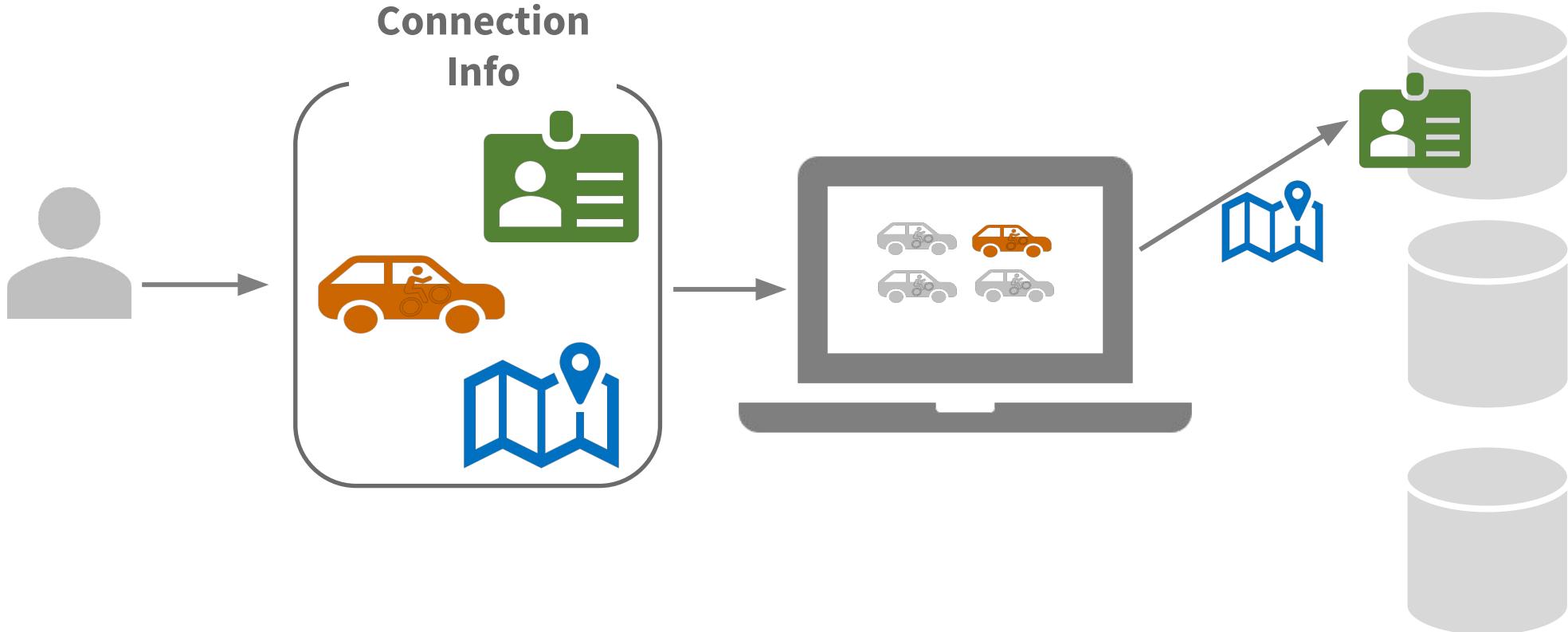


- URL
 - IP Address
-

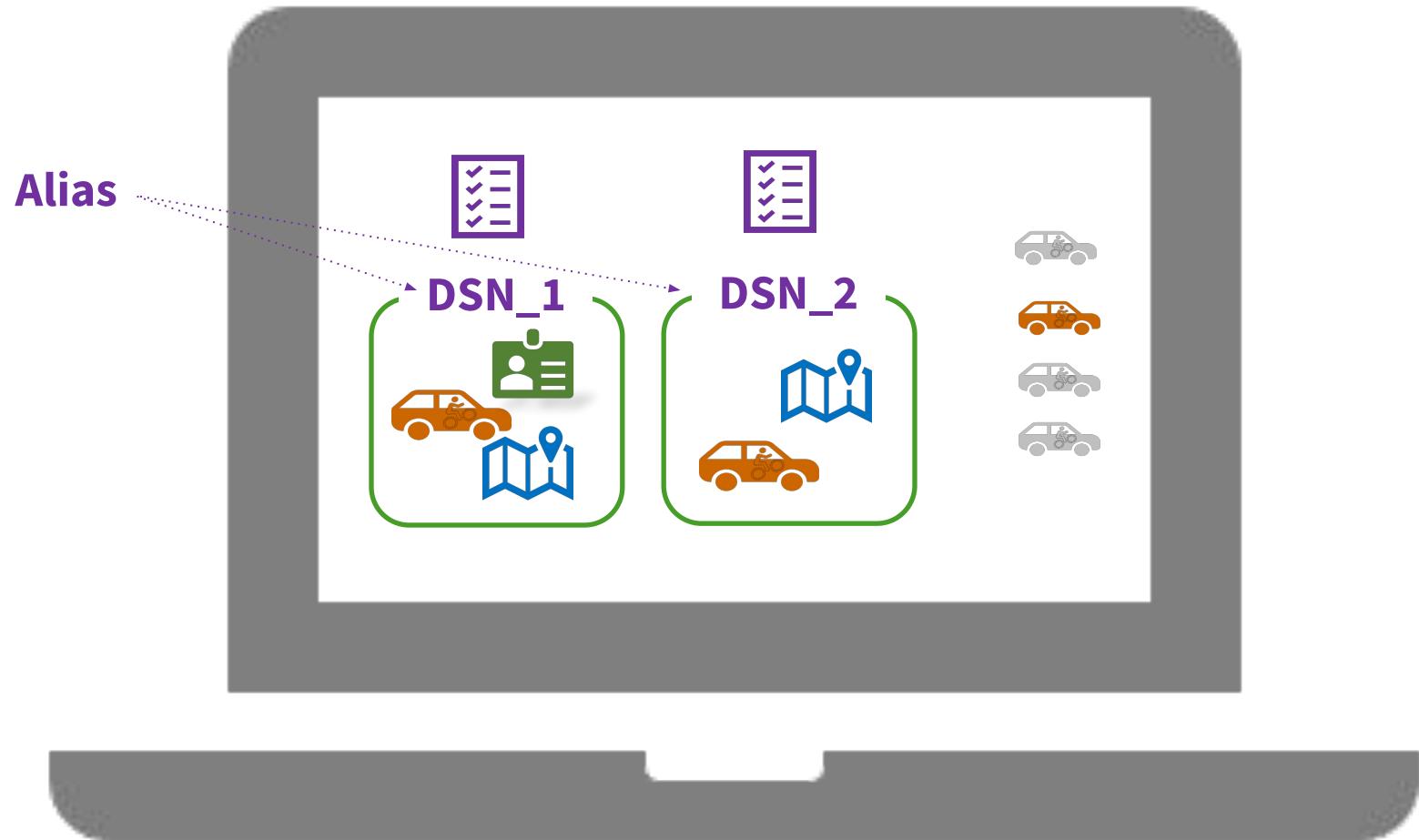


- ODBC (Used by **ADO & OLE DB**)
- JDBC

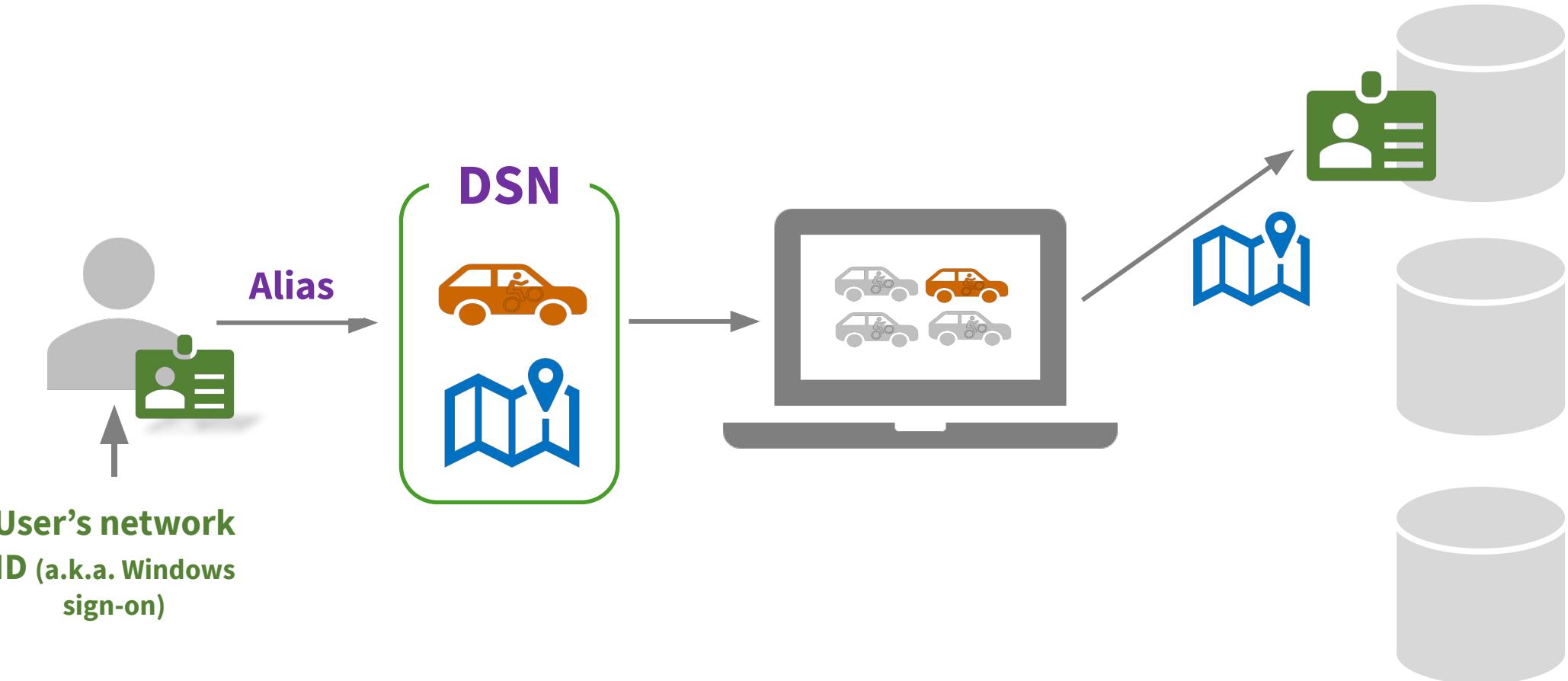
Connection info



Data Source Name (DSN)



The ideal connection



The connections pane

The screenshot shows the RStudio interface with the Connections pane highlighted by a blue rounded rectangle. The Connections tab is selected in the top navigation bar. A connection named "postgres - rstudio_dev@localhost" is listed. Below the pane, the RStudio file browser displays a directory structure for "big-data".

Code Editor:

```
1 ````{r db-connections, include = FALSE}
2 if(Sys.getenv("GLOBAL_EVAL") != "") eval_connections <-
  Sys.getenv("GLOBAL_EVAL")
3 eval_connections <- FALSE
4
5
6 ````{r, eval = eval_connections, include = FALSE}
7 library(DBI)
8 library(odbc)
9 library(config)
10 library(keyring)
11
12
13 # Introduction to database connections
14
15 ## Connect with the Connections pane
```

File Browser:

Name	Size	Last Modified
..		
.gitignore	100 B	Jan 1
.Rbuildignore	28 B	Jan 1
.Renviron	40 B	Jan 1
01-intro-to-vroom.Rmd	4.7 KB	Jan 1
02-intro-to-dtplyr.Rmd	4.8 KB	Jan 1
03-db-connections.Rmd	4.3 KB	Jan 1
04-intro-to-DBI.Rmd	4.8 KB	Jan 1
05-db-analysis.Rmd	3.4 KB	Jan 1

Connections Pane:

New Connection

Connection Status

postgres - rstudio_dev@localhost

Connections Sidebar:

- Introduction to da...
- Connect with the ...
- Connecting via D...
- Connect with a c...
- Secure connectio...

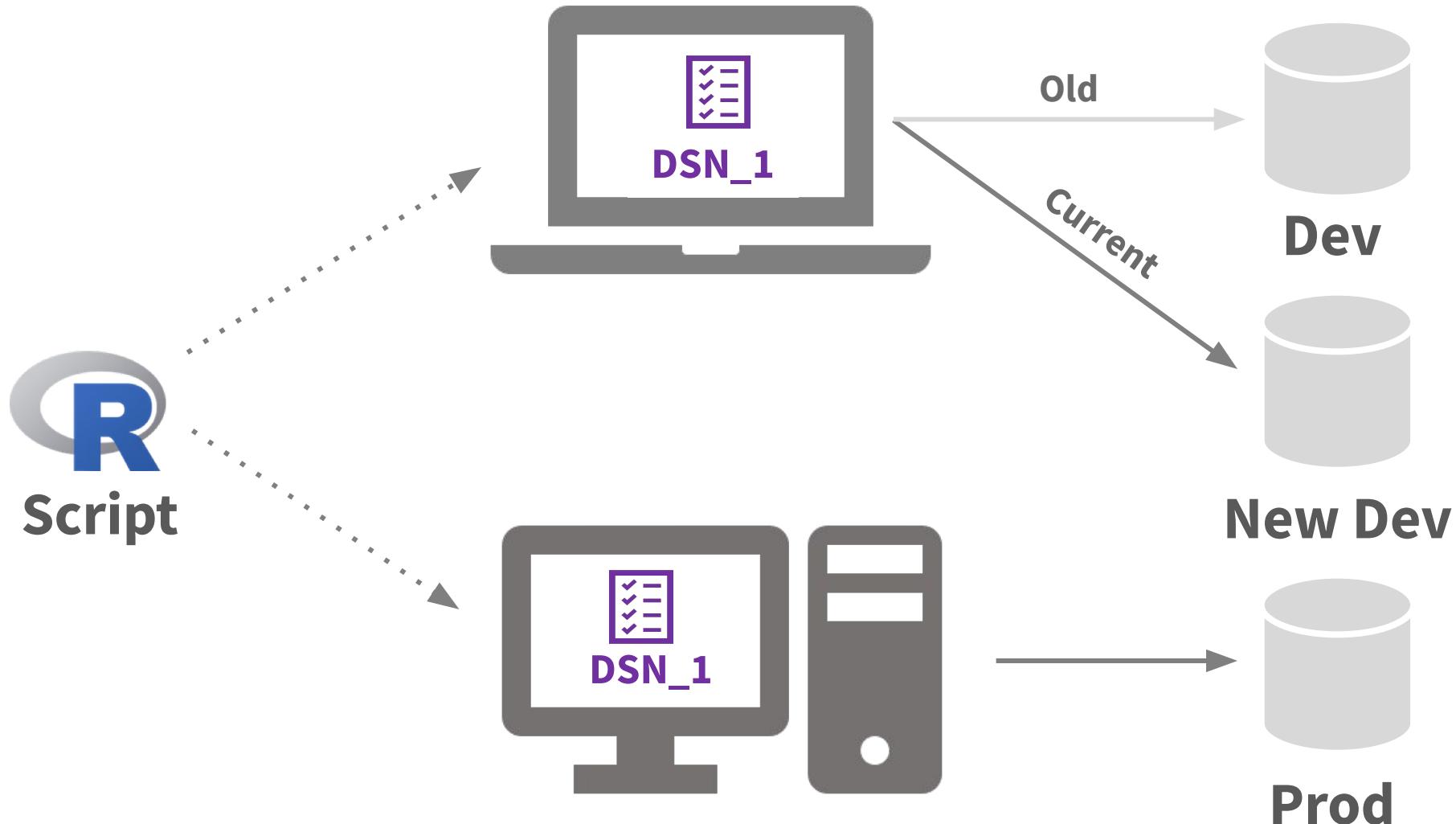
New Connection

Connect to Existing Data Sources

- Postgres Dev
- Postgres Prod
- SQL Server (DSN)
- Pins
- Livy
- Spark
- Athena
- BigQuery

Using RStudio Connections

Why DSN?



Alternatives for securing connections

1. config
2. keyring
3. Environment variables
4. options()
5. Prompt for credentials

R packages

General connections

- DBI
- odbc
- connections

Specific Connections

- sparklyr
- RPostgres
- RSQLite
- . . .

ODBC Drivers



Athena



BigQuery



Cassandra



Hive



Impala



MongoDB



MySQL



Netezza

ORACLE



PostgreSQL



Redshift

TERADATA

Teradata



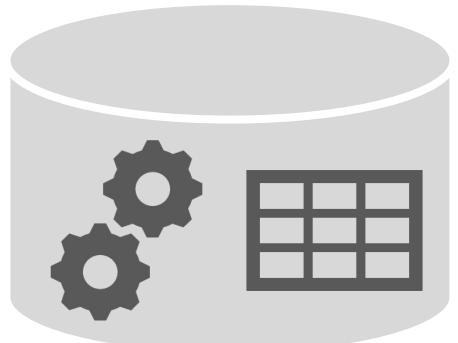
Salesforce



SQL Server

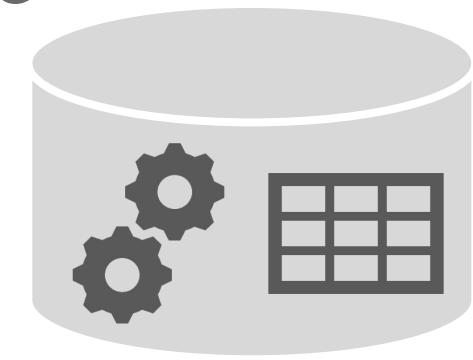
Wrangle inside the DB

Time Consuming



Extract Data

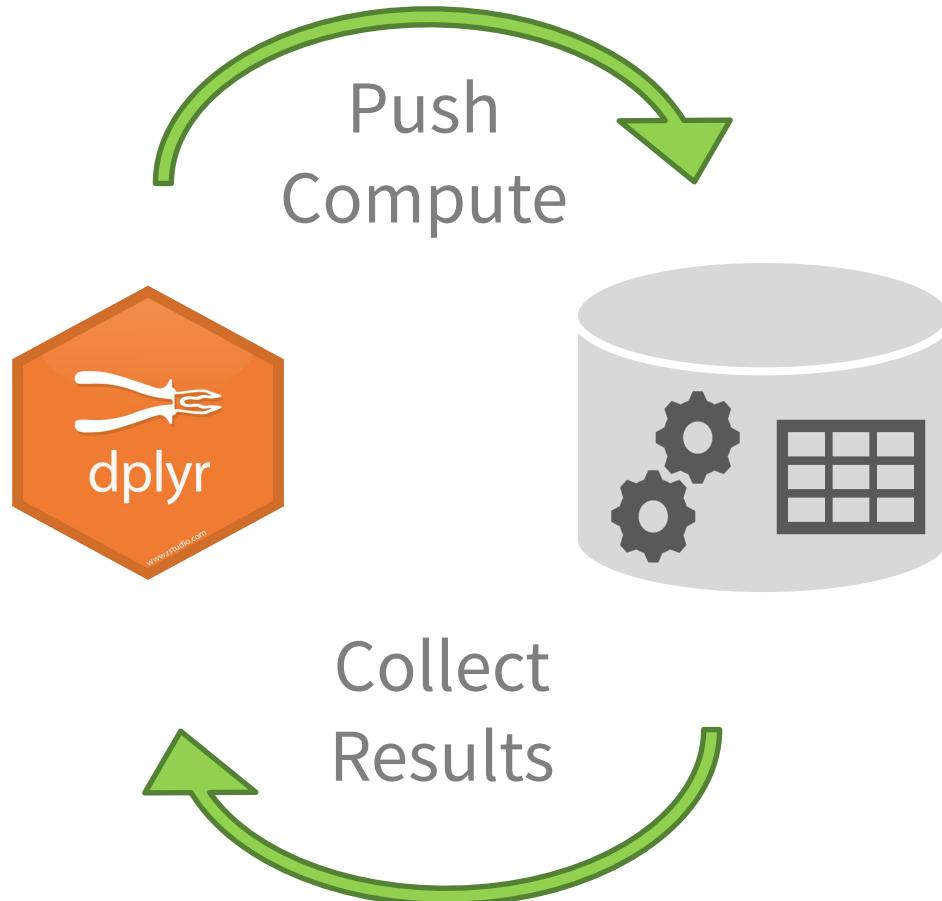
Push
Compute



Collect
Results



Advantages



1. dplyr translates to SQL
2. Take advantage consistent syntax
3. All your code is in R!

DBI package

1. Stands for **database interface**
2. Helps connect R to various database management systems
3. Used for connecting to and interacting with various databases
4. Execute SQL commands against the database



DBI common functions

Connecting

- dbConnect
- dbDisconnect

Queries

- dbSendQuery
- dbGetQuery
- dbExecute

Tables

- dbListTables
- dbWriteTable
- dbReadTable

Options to Push Compute

Write SQL statements

```
SELECT "customer_id",
COUNT(*) AS "n"
FROM "retail.orders"
GROUP BY "customer_id"
```

Use dplyr verbs

```
orders %>%
  count(customer_id)
```

sparklyr package

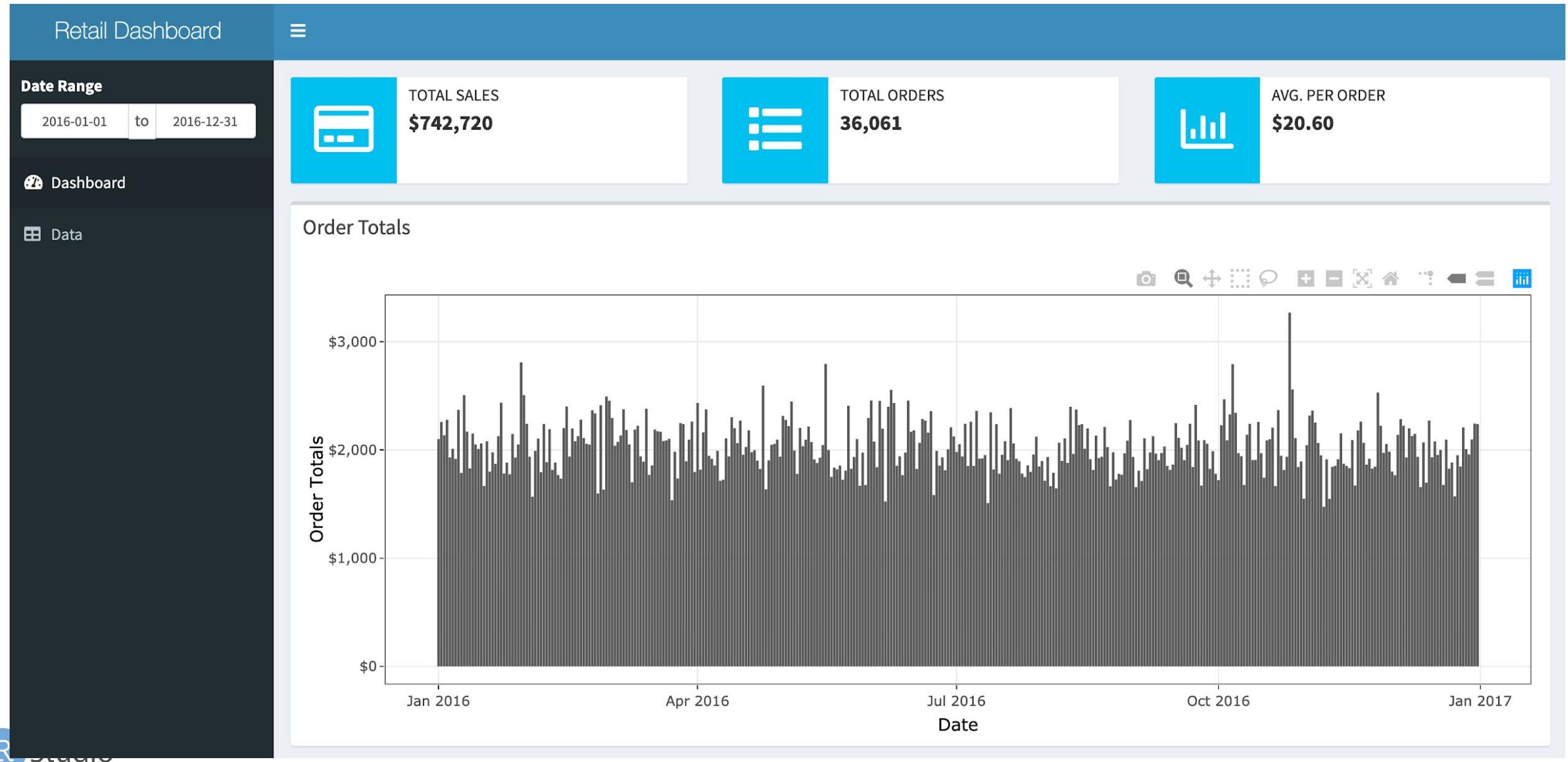
1. Leverage the power of Spark from R
2. Take advantage of distributed,
in-memory processing
3. Create extension that leverage the
Spark API



Best Practices

- Use `dplyr` for consistent syntax and tooling
- Rely on each tool for its strengths
- Leave heavy computation to tools designed for heavy computation (Databases, Spark)
- Use R for final analysis
- When possible, avoid bringing **all** data into R
- Create extracts and views to speed things up

One more thing...



Bookmark and check regularly

- <http://db.rstudio.com/>
- <http://spark.rstudio.com/>
- <https://www.tidyverse.org/>
- <https://rviews.rstudio.com/>
- <https://rviews.rstudio.com/categories/databases>
- <https://blog.rstudio.com/>

Join the community!

R Studio Community

all categories ► all tags ► Categories Latest New (12) Unread Top

Category	Topics	Latest
 rstudio::conf 2018 This category is for anything and everything related to rstudio::conf.	4 / week 2 new	 How can I connect R with v application • new rstudio
 tidyverse This category is for anything and everything about the tidyverse.	23 / week	 □ Crash when quitting ■ RStudio IDE bug
 RStudio IDE This category is for discussing the RStudio IDE, both	16 / week 3 new	 □ Is there a way to measure • new

<https://community.rstudio.com/>

Familiarize yourself with the repos

If I need to...	Check out
Report an issue or see if others are having the same problem	Issues
See if an feature exists or if it's coming up in future releases	NEWS
See the basics about the package	README

- <https://github.com/tidyverse/dplyr>
- <https://github.com/tidyverse/dbplyr>
- <https://github.com/tidyverse/ggplot2>
- <https://github.com/r-dbi/odbc>
- <https://github.com/r-dbi/DBI>
- <https://github.com/edgararuiz/dbplot>
- <https://github.com/tidymodels/tidypredict>
- <https://github.com/rstudio/sparklyr>



Thank you