Presentation Slides | Video recording

Big Data with Modern R & Spark from **Xavier de Pedro**[1]

Big data with Modern R & Spark in Public Statistics

"SEMINAR: *Análisis de Big data con Tidyverse y Spark: uso en estadística pública*"
By Xavier de Pedro Puente, Ph.D.
Technician at the Municipal Data Office. Barcelona City Council.

Monday September 14, 2020. 16h-17:30h + questions
Within the context of the postgraduate course on
"**Data Science. Applications to Biology and Medicine with Python and R**"
at University of Barcelona. 2020.

IImage from https://blog.rstudio.com/2019/03/15/sparklyr-1-0/[2]

Outline

1. About me
2. Barcelona City Council Çase
3. Modern R: Tidyverse
4. Modern R & Big Data
5. Spark (Apache) & Sparklyr (RStudio)
6. Alternative approach

(1) About me

Xavier de Pedro Puente, Ph.D. xavier.depedro (a) seeds4c.org

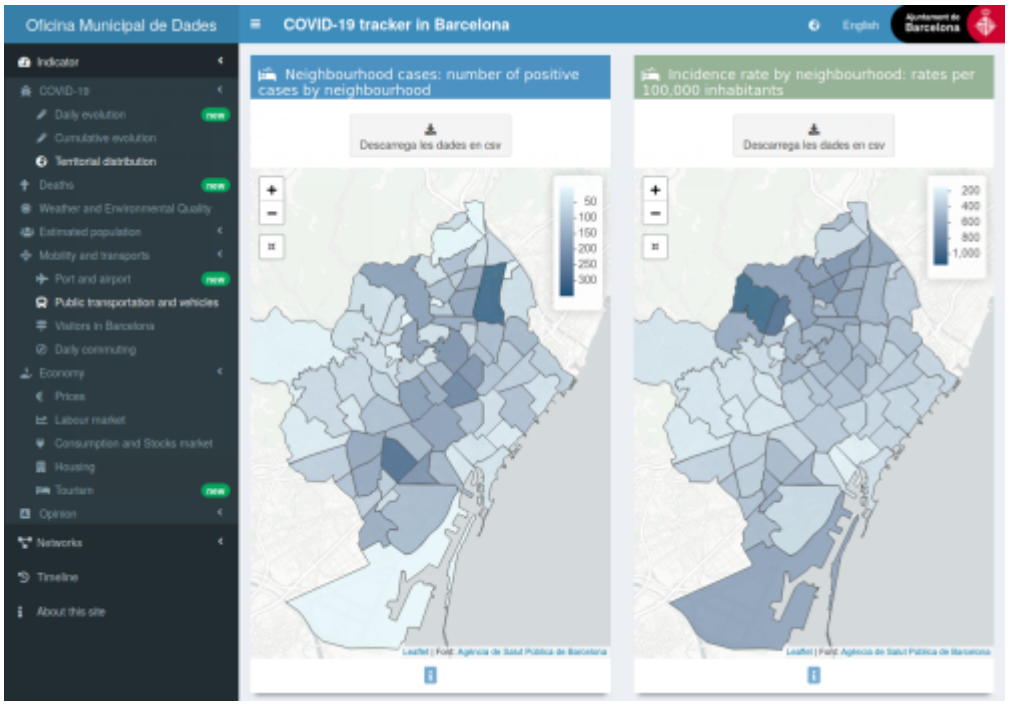
- **Academics**:
  - Degree in Biology (UB[3])
  - Ph.D. in Ecology (UB[4])
  - Postgraduate in Bioinformatics (UOC[5])

- **Current Work**:
  - Senior technician at Municipal Data Office (OMD[6], Barcelona City Council[7])
- **Past (related) Work**:
  - Bioinformatics technician (UEB[8], VHIR[9])
  - Systems administrator (UEB[10], VHIR[11])

Disclaimer The views expressed in this presentation are those of the author and do not necessarily represent the views and policies of the Barcelona City Council. Any mention of trade names, products or services does not imply an endorsement by the Barcelona City Council unless explicitly stated as such.

(2) Barcelona City Council Case

- 2018: OMD - Municipal Data Office opened[12] (details[13]):
  Management, quality, governance and use of data controlled and/or stored by Barcelona City Council and all of its associated bodies (both public and private).
  .
- Public Statistics[14] portal
- Recent product[15]: Covid-19 in BCN monitoring with an R Shiny App[16]



Barcelona City Council: Public money - public code

- Barcelona: first city to join the Free Software Foundation, Public Money, Public Code[17] campaign

- One of the case studies[18] of the use of open-source software and open code to democratise cities.

There is a Government policy for open source and agile methodologies[19]

Barcelona City Council: CityOS

City Council's infrastructure based on open-source Big Data technology

Video[21]

City OS: internal data management, known as "Data Lake"[20]

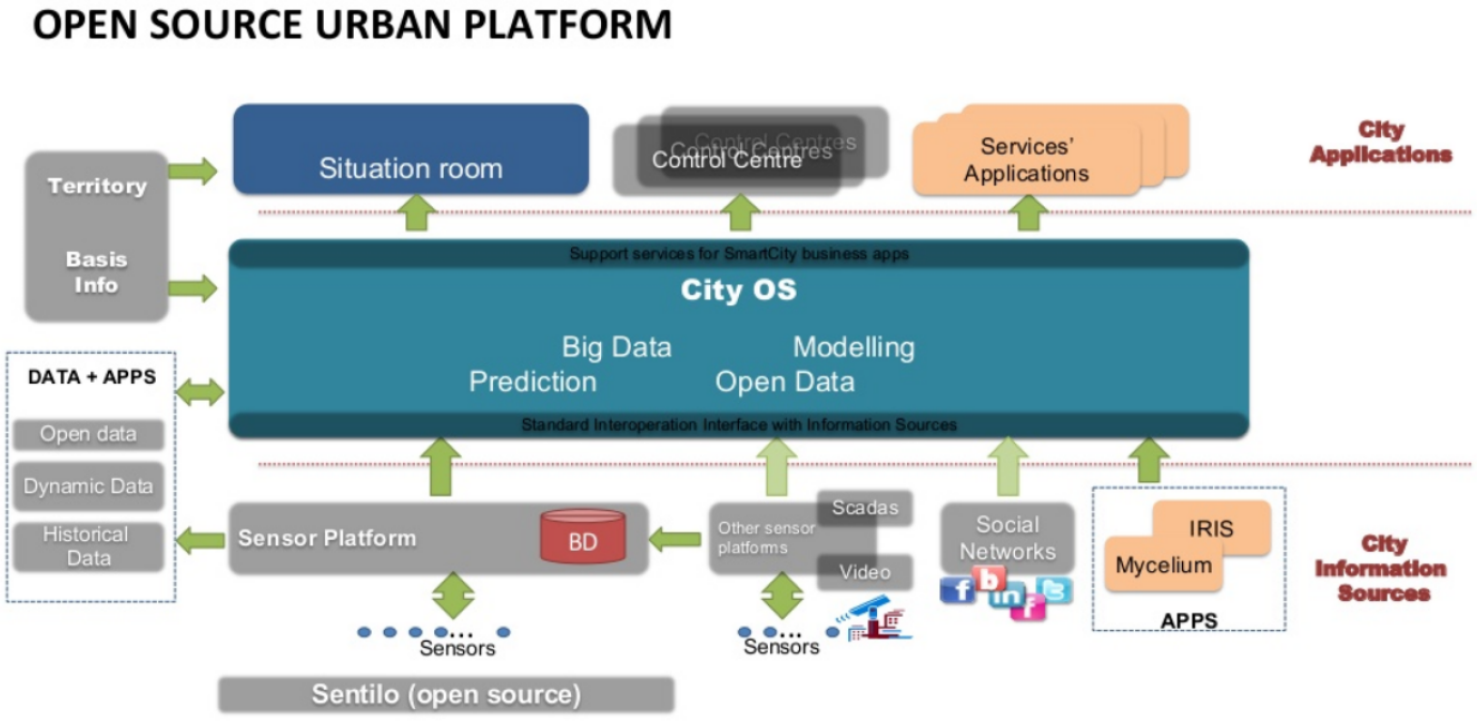
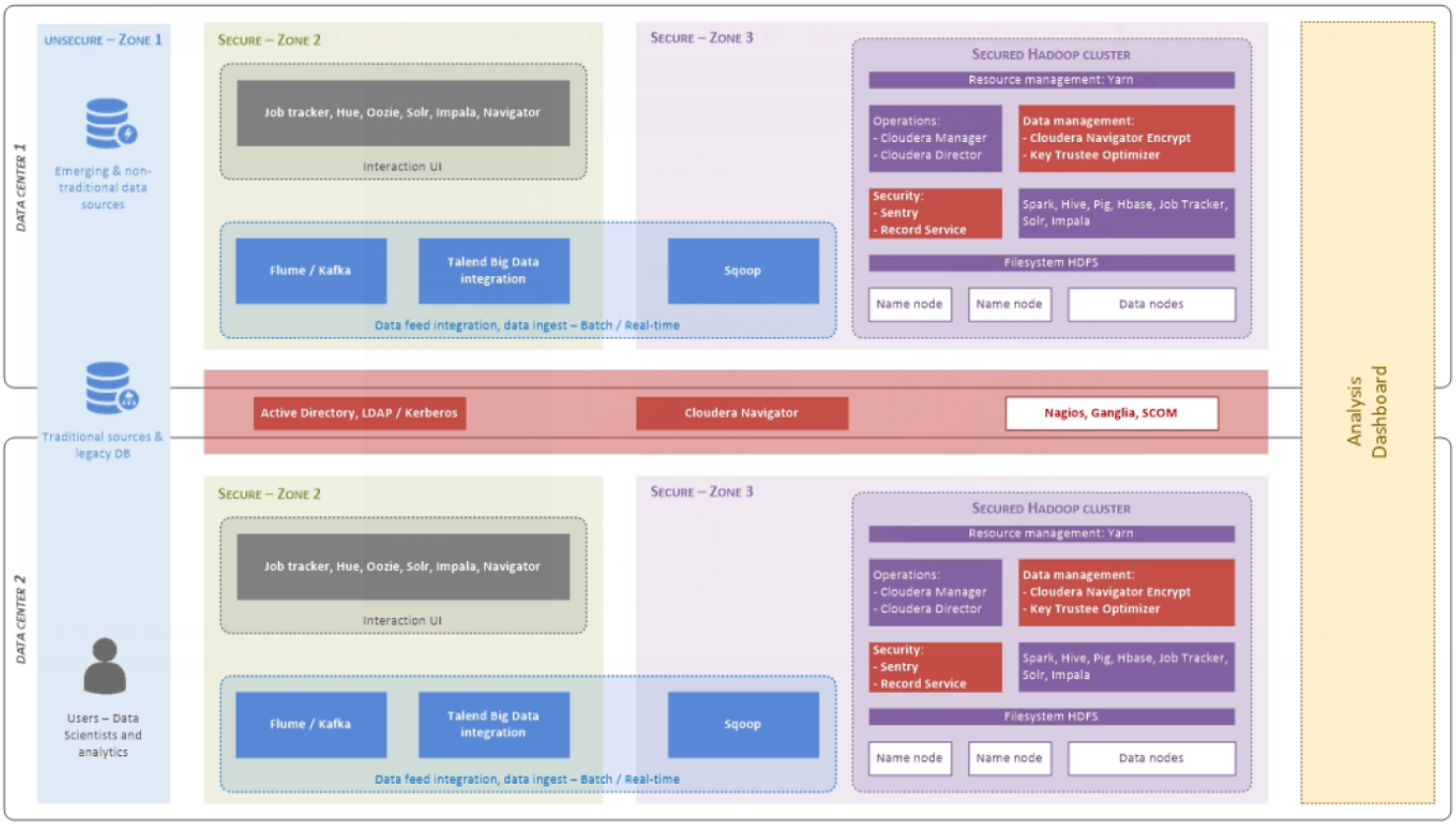


Image from Francesca Bria[22] (Barcelona Digital City Roadmap 2017-2020)

Barcelona City Council: CityOS Technologies

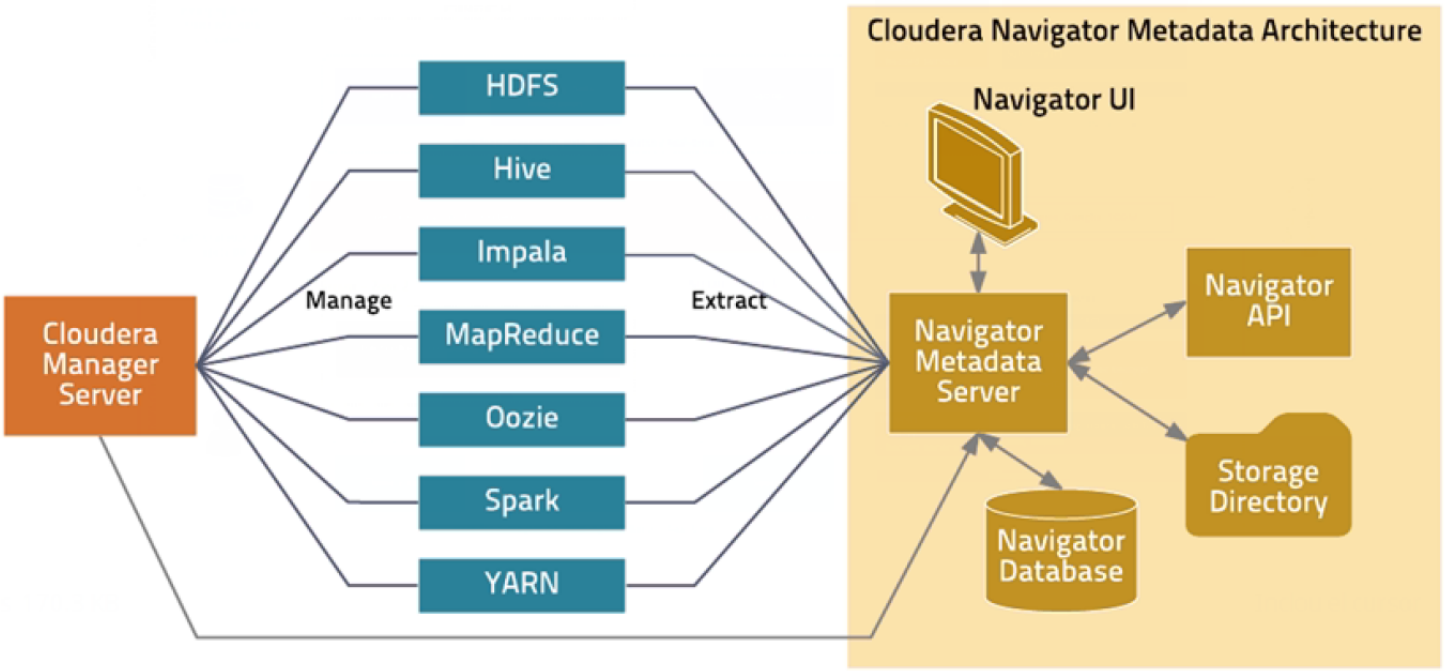
CityOS technologies: GNU/Linux, CentOS, Cloudera, Activiti, Talend, Protégé, **R**, Zabbix, Nagios, Ganglia



Source: https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN[23] - Image from J. Berdonces a Github[24]

Barcelona City Council: CityOS - Cloudera Manager

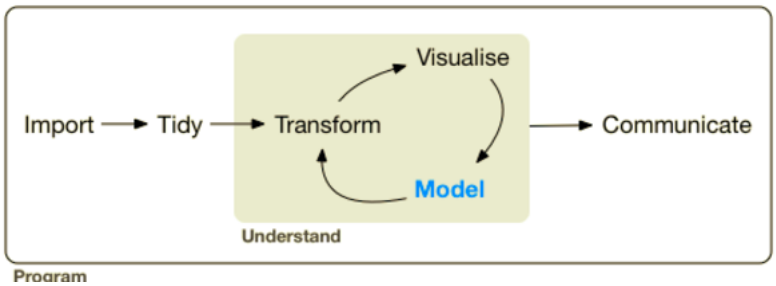
Cloudera with Hadoop File System, Hue, HBase, Hive, Impala, Oozie, **Spark**, Yarn, Kafka, ...



Source: https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN[25] - Image from J. Berdonces a Github[26]

(3) From Base R to "Modern R": Tidyverse

- 2000: First stable beta version (v1.0) released.
- ... "Base R" ...
- 2016: Tidyverse[27] 1.0.0 package released on CRAN[28]
- 2016: "Introduction to Modern R - r4stats.com"[29]
- 2017: "Martin Hadley on R and the Modern R Ecosystem - InfoQ Podcasts"[30]
- 2017: "Modern Data Science with R - Chapman & Hall/CRC"[31]
- 2019: "Modern R with the tidyverse - Econometrics and Free Software"[32] (blog)
- 2019: "Modern R and the Tidyverse - Data Science Workshops"[33]
- 2020: "Modern R: Welcome to the tidyverse"[34] (video tutorial)
- 2020: Statistical Inference via Data Science: ModernDive into R & Tidyverse"[35]



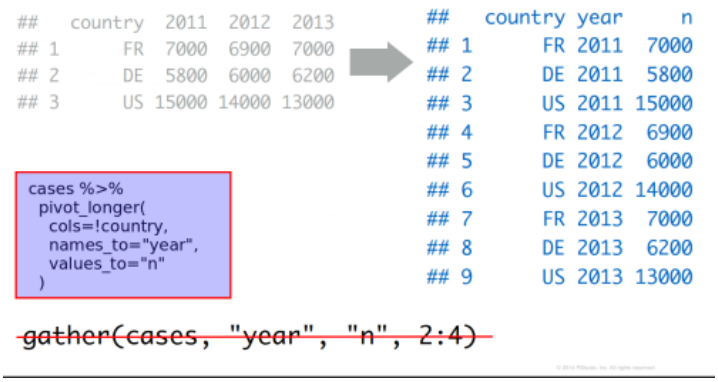
Modern R (Tidyverse)

Workflow, Packages, People/Community

Derived from here[36] & here[37]

Modern R (Tidyverse) principles

1. **Main structures are ordered data**
   - Each variable is saved in its own column.
   - Each observation is saved in its own row.
   - Each "type" of observation stored in a single table
2. .**Each function represents one step**
3. **Functions are combined with the pipe operator** %>%
4. **Each step is a query or command**



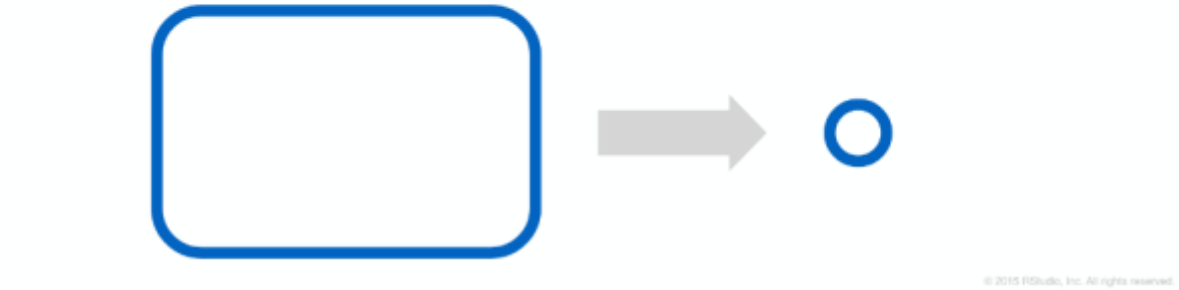
Derived from here[38], here[39] & here[40]

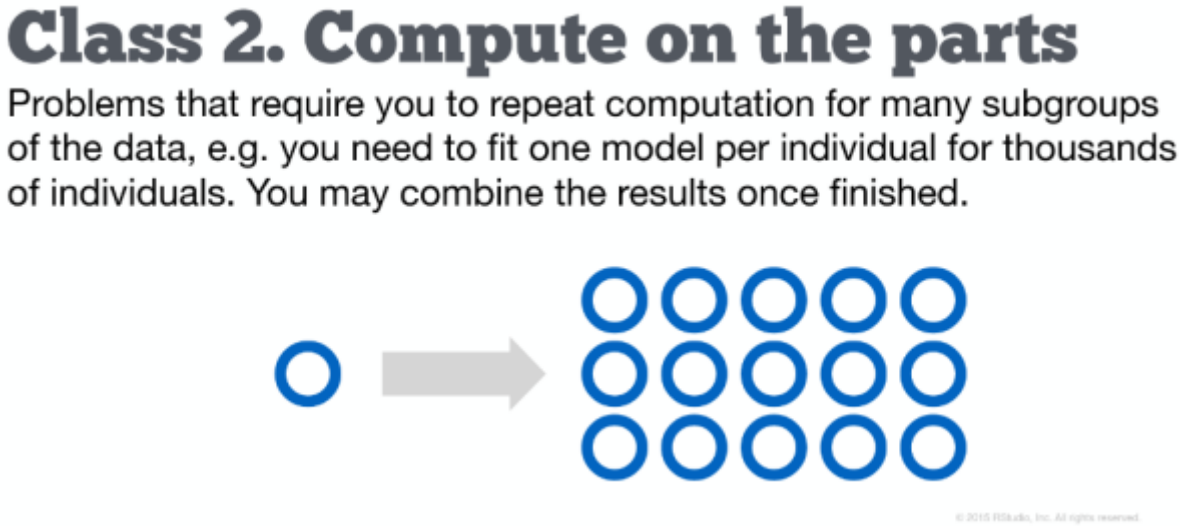
(4) Modern R & Big Data

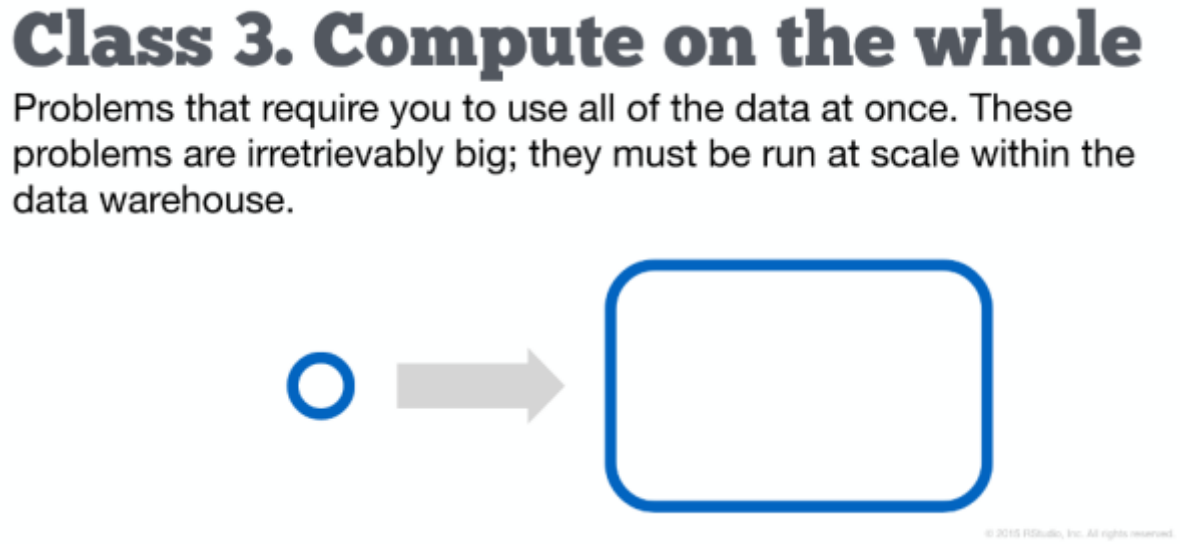
Data > RAM = Problem

From an RStudio seminar[41] by Garrett Grolemund[42]

3 classes of Big Data Problems

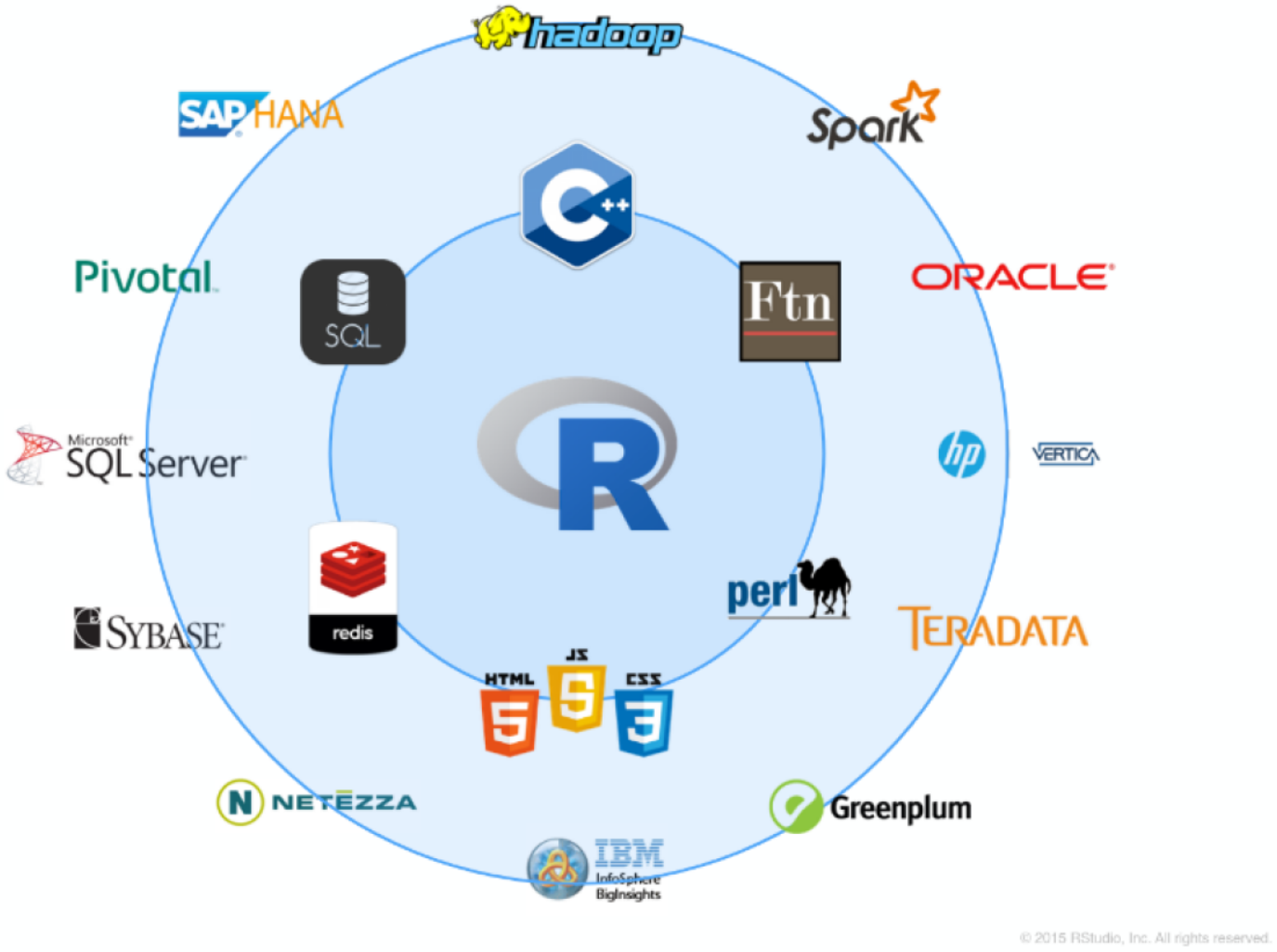






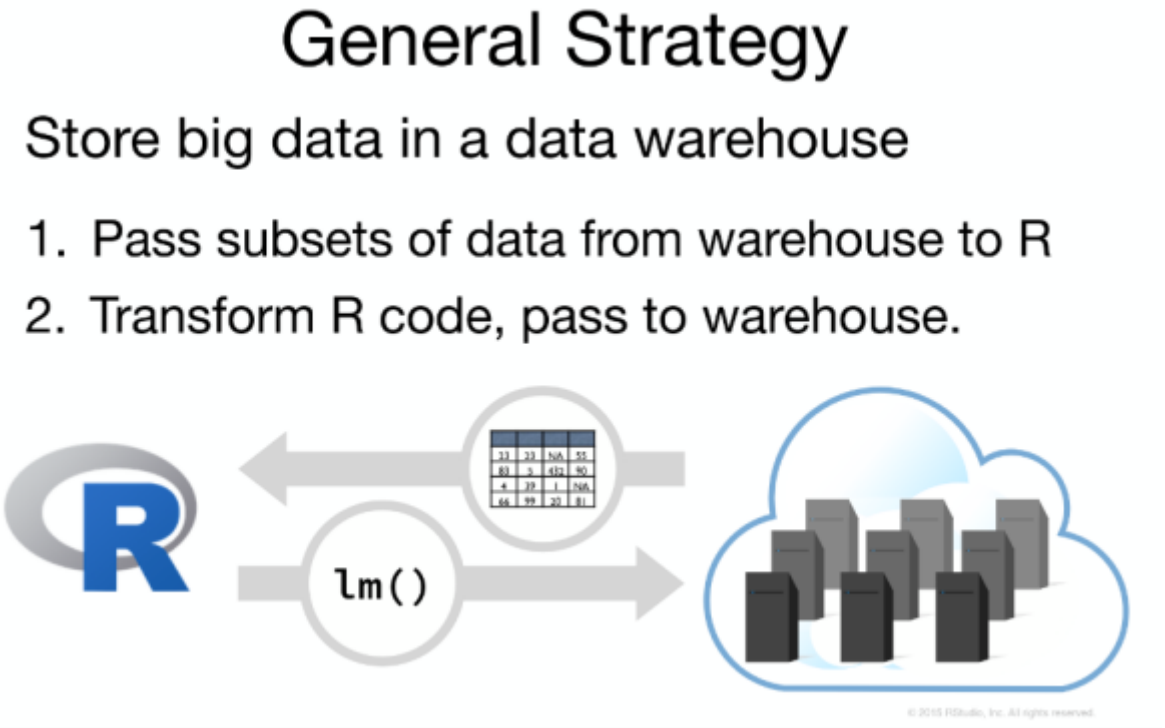
From an RStudio seminar[43] by Garrett Grolemund[44]

R interacts with other technologies



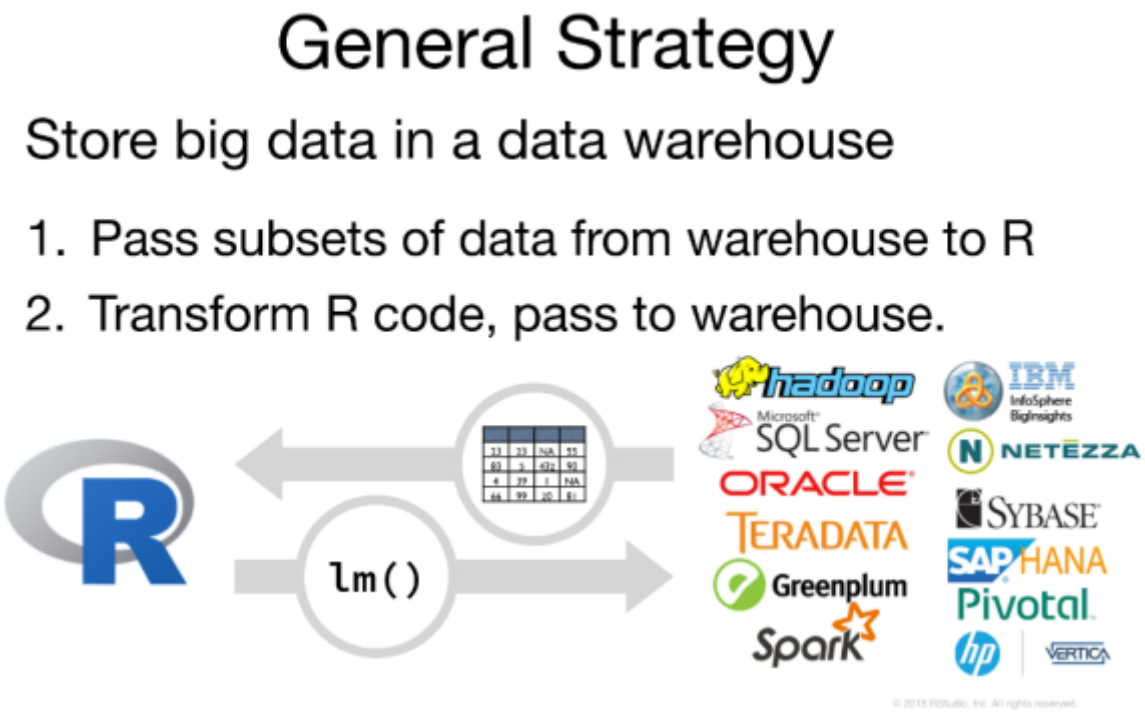
From an RStudio seminar[45] by Garrett Grolemund[46]

R General Strategy (i): Class 1 & 2



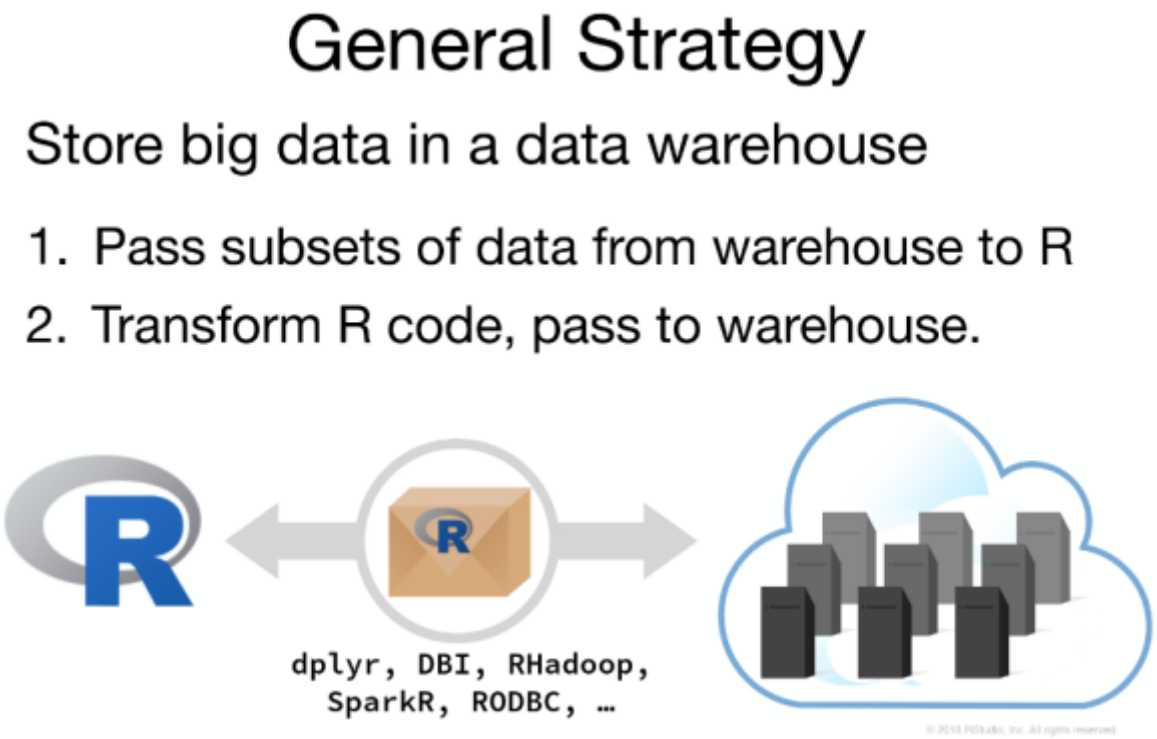
From an RStudio seminar[47] by Garrett Grolemund[48]

R General Strategy (ii): Class 1 & 2



From an RStudio seminar[49] by Garrett Grolemund[50]

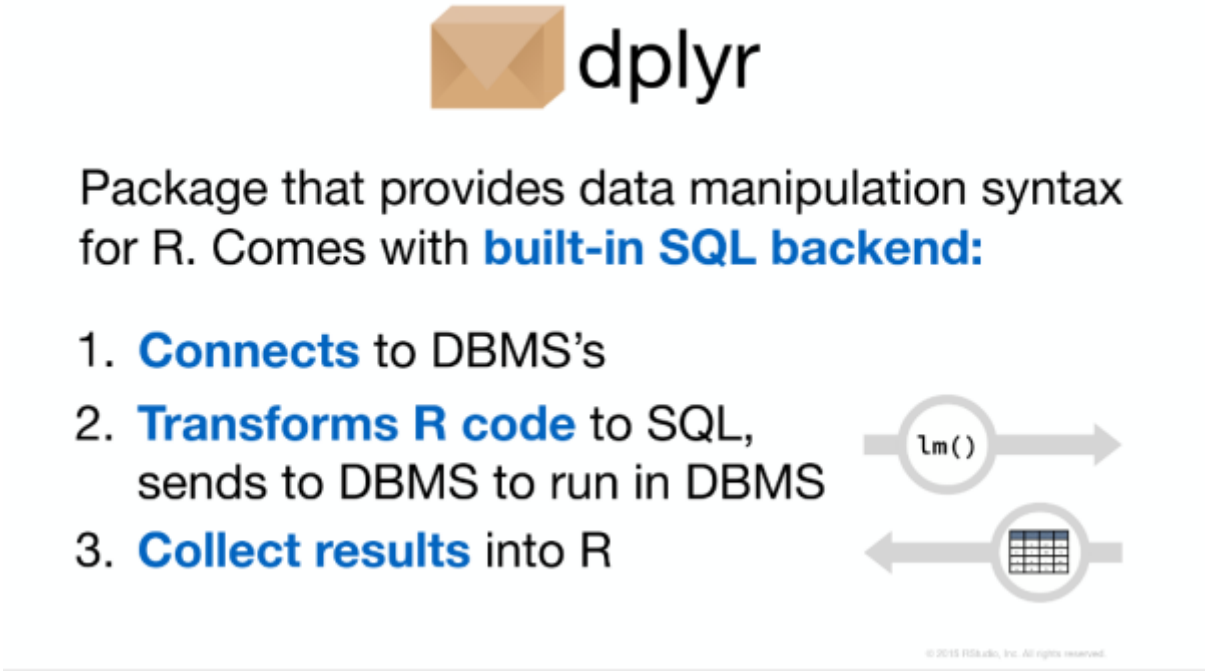
R General Strategy (iii): Class 3



From an RStudio seminar[51] by Garrett Grolemund[52]

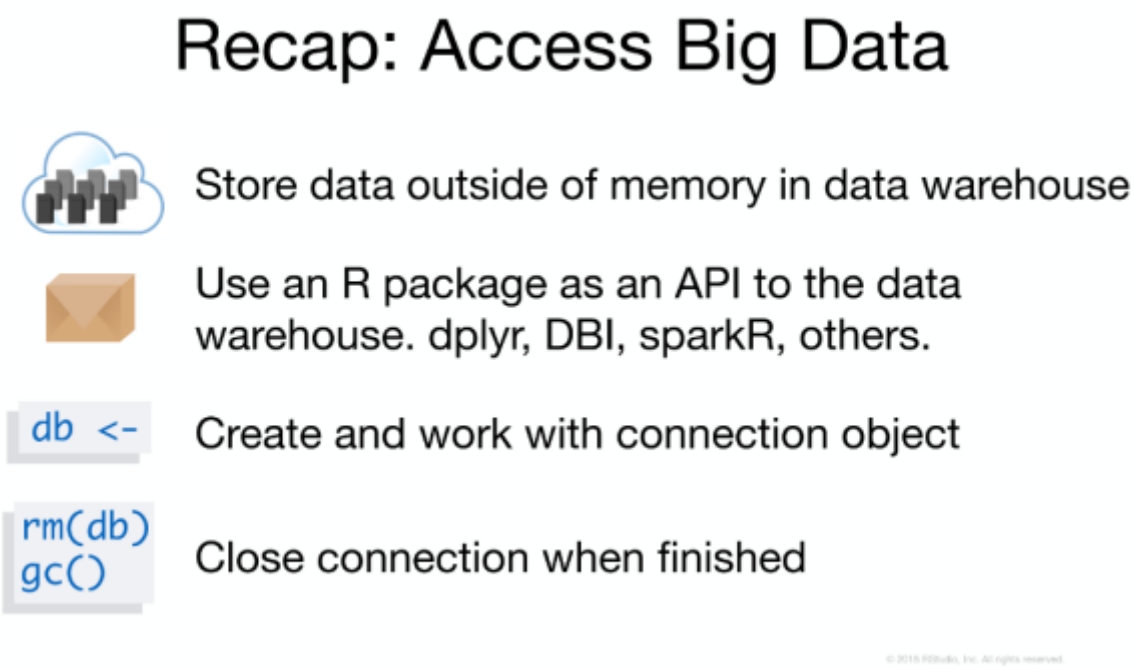
Class 3 - dbplyr

R - tidyverse - dplyr + dbplyr



From an RStudio seminar[53] by Garrett Grolemund[54]

R General Strategy (iv): Recap



From an RStudio seminar[55] by Garrett Grolemund[56]

(5) Spark (Apache) & sparklyr (Rstudio)

- Spark[57], from Apache Foundation[58]: a leading tool that is democratizing our ability to process large datasets.
- R Packages:
  - Apache introduced an interface for the R computing language: **SparkR** R package
  - RStudio introduced **sparklyr** R package: a project merging R and Spark into a powerful

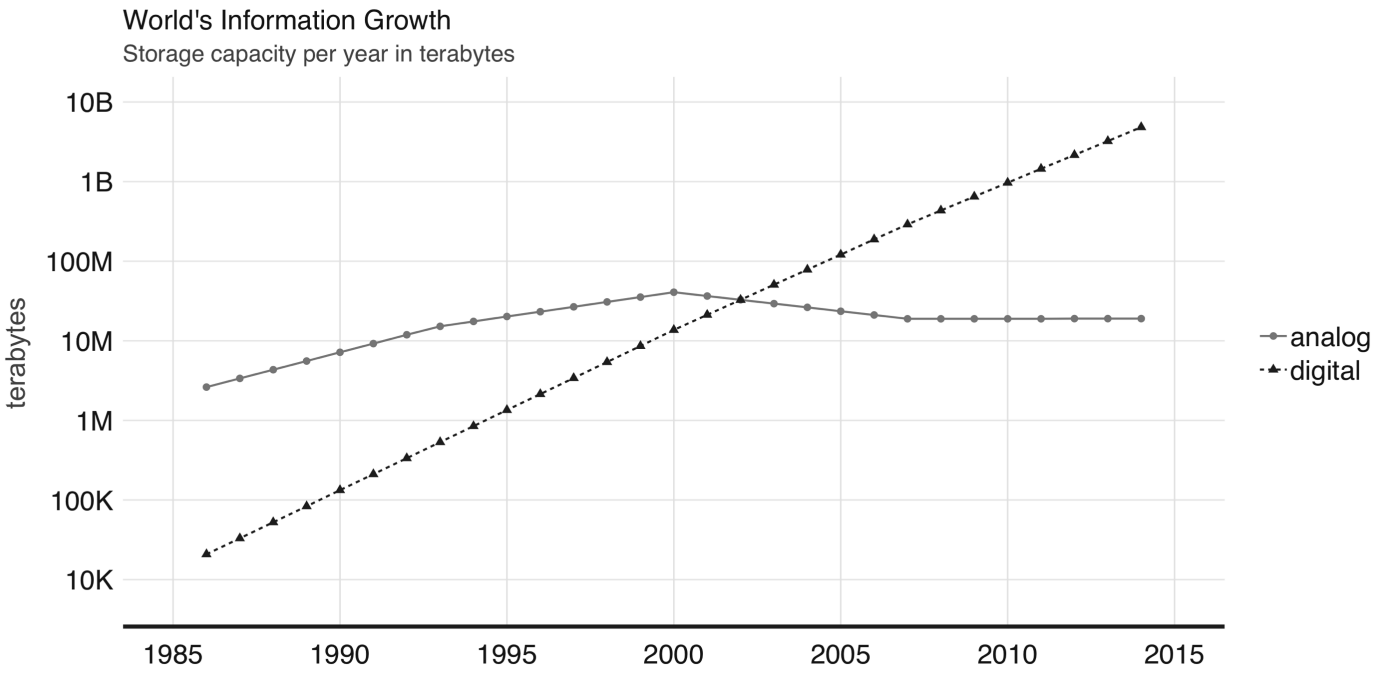
tool that is easily accessible to all.

- But **why where they introduced**?
- And **why 2 different packages**?

Derived from here[59] & here[60]

Digital information vs analog information | World Bank - 2003

- World Bank report: digital information surpassed analog information around 2003.
  - 10 million terabytes of digital information (~ 10 million storage drives today)
  - our footprint of digital information is growing at exponential rates.



From "The R In Spark" (book)[61]

Google File System | Google - 2003

- Search engines were unable to store all of the web page information required to support web searches in a single computer.
- They had to split information into several files and store them across many machines
- This approach became known as "**Google File System**", due to a research paper published in 2003 by Google.

From "The R In Spark" (book)[62]

MapReduce | Google - 2004

- 2004: Google published a new paper describing how to perform operations across the Google File System:
  - approach called "MapReduce"
    - map operation: arbitrary way to transform each file into a new file
    - reduce operation: combines two files
  - Both operations require custom computer code, but the MapReduce framework takes care of automatically executing them across many computers at once.
  - Sufficient to process all the data available on the web, while providing flexibility to extract meaningful information from it.

From "The R In Spark" (book)[63]

Hadoop Distributed File System (HDFS) | Yahoo - 2006

- A team at Yahoo implemented the Google File System and MapReduce as a single open source project, released in 2006 as **Hadoop**
  - Google File System implemented as the Hadoop Distributed File System (**HDFS**).
- The Hadoop project made distributed file-based computing accessible to a wider range of users and organizations, making MapReduce useful beyond web data processing.

From "The R In Spark" (book)[64] | Image from De Apache Software Foundation, with Apache License 2.0[65]

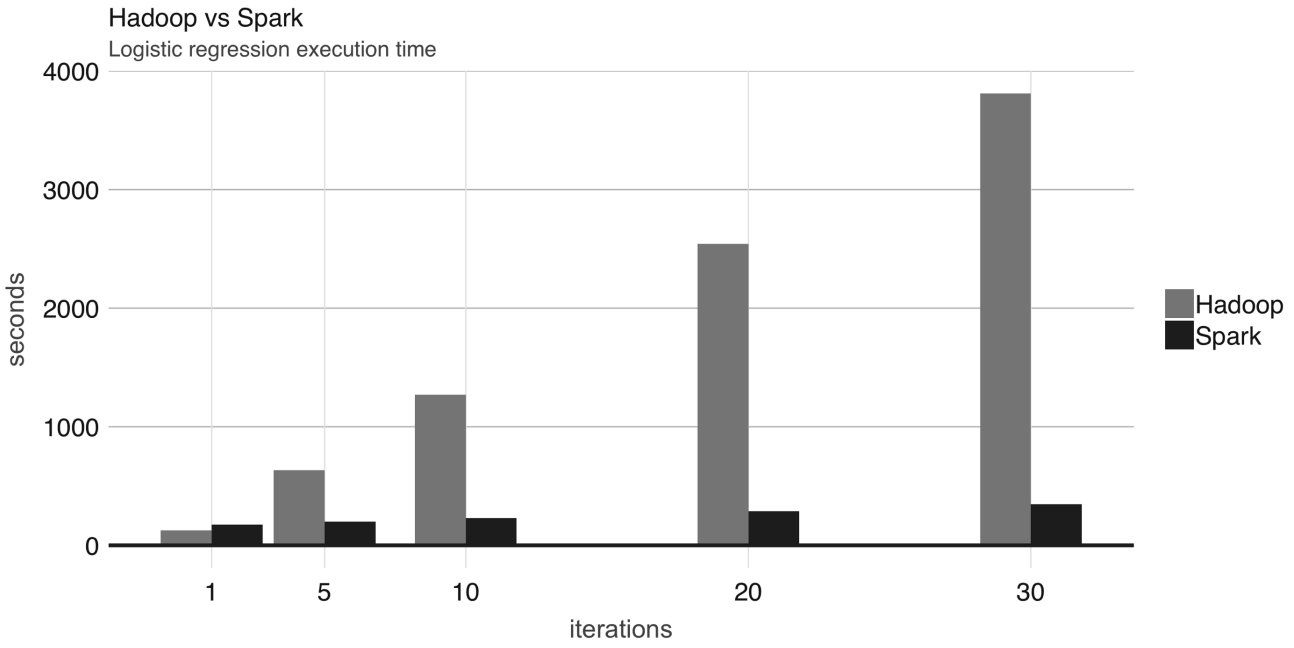
Hive | Facebook - 2008

- Hadoop provided support to perform MapReduce operations over a distributed file system, but it still required MapReduce operations to be written with code every time a data analysis was run
- **Hive** project (2008, by Facebook) brought **Structured Query Language** (SQL) support to Hadoop.
  - Data analysis could now be performed at large scale without the need to write code for each MapReduce operation

From "The R In Spark" (book)[66] | Image from Davod[67] with Apache License 2.0[68]

Spark (closed sourced) | UCBerkely - 2009

- In 2009, **Spark** began as a research project at UC Berkeley's AMPLab to improve on MapReduce.
  - Spark provided a richer set of verbs beyond MapReduce to facilitate optimizing code running in multiple machines.
  - **Spark also loaded data in-memory**, making operations much faster than **Hadoop's on-disk storage**.



From "The R In Spark" (book)[69]

Spark: In-memory and on-disk

- Spark: well known for its in-memory performance, but designed to be a general execution engine that works both in-memory and on-disk
  - For instance, Spark has set sorting, for which data was not loaded in-memory, but using improvements in network serialization, network shuffling, and efficient use of the CPU's cache to dramatically enhance performance.
  - If you needed to sort large amounts of data, there was no other system in the world faster than Spark.

From "The R In Spark" (book)[70]

Spark: faster and easier

- Spark is much faster, more efficient, and easier to use than Hadoop.
- Speed Example:
  - Without Spark: it takes 72 minutes and 2,100 computers to sort 100 terabytes of data using Hadoop,
  - With Spark: only 23 minutes and 206 computers

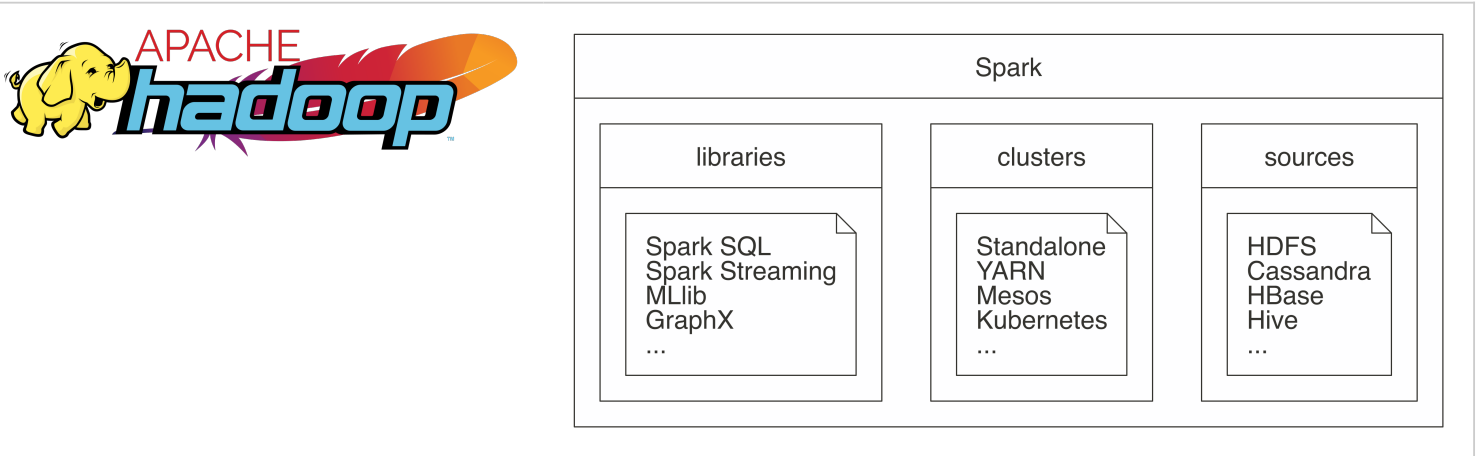
- Simplicity example: word-counting MapReduce example takes:
  - about 50 lines of code in Hadoop, but
  - only 2 lines of code in Spark.

	Hadoop Record	Spark Record
Data Size	102.5 TB	100 TB
Elapsed Time	72 mins	23 mins
Nodes	2100	206
Cores	50400	6592
Disk	3150 GB/s	618 GB/s
Network	10Gbps	10Gbps
Sort rate	1.42 TB/min	4.27 TB/min
Sort rate / node	0.67 GB/min	20.7 GB/min

From "The R In Spark" (book)[71]

Spark (open sourced) - 2010 | Apache Foundation - 2013

- 2010: open sourced. 2013: donated to the Apache Software Foundation
- Apache Spark is a unified analytics engine for large-scale data processing
  - Unified: supports many libraries, cluster technologies, and storage systems.
  - Analytics: discovery & interpretation of data to communicate information
  - Engine: expected to be efficient and generic.
  - Large-Scale: as cluster-scale (set of connected computers working together).



From "The R In Spark" (book)[72] | Image from the Apache Software Foundation, with Apache License 2.0[73]

SparkR (Base R) vs. sparklyr (Modern R)

Feature	SparkR	sparklyr
Data input & output	+ +	+ +
Data manipulation	-	+ + +
Documentation	+ +	+ +
Ease of setup	+ +	+ +
Function naming	- -	+ + +
Installation	+	+ +
Machine learning	+	+ +
Range of functions	+ + +	+ +
Running arbitrary code	+	+ +
Tidyverse compatability	- - -	+ + +

From https://www.eddjberry.com/post/2017-12-05-sparkr-vs-sparklyr/[74]

Sparklyr | RStudio

- Sparklyr, from Rstudio: https://spark.rstudio.com/[75]
- R interface for Apache Spark, agnostic to Spark versions,
  - 2016: 1st version released (v0.4)
- Easy to install, serving the R community,
- Embracing other packages & practices from the R community (tidyverse, ...)
- Designed for: New Users, Data Scientists, and Expert Users

Image from here[76]

Summary: Big Data with Modern R & Spark

Big Data with Modern R & Spark in context

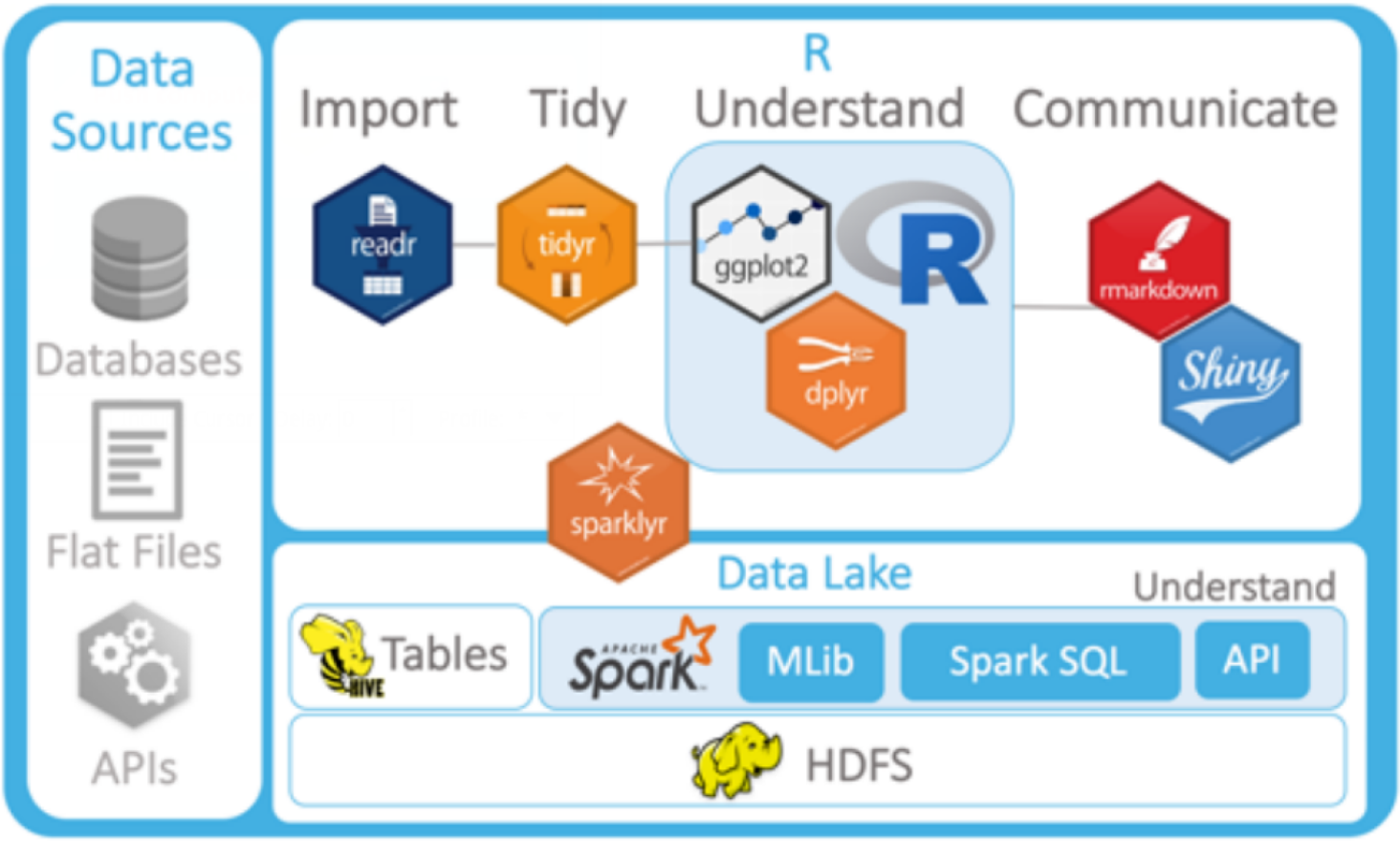
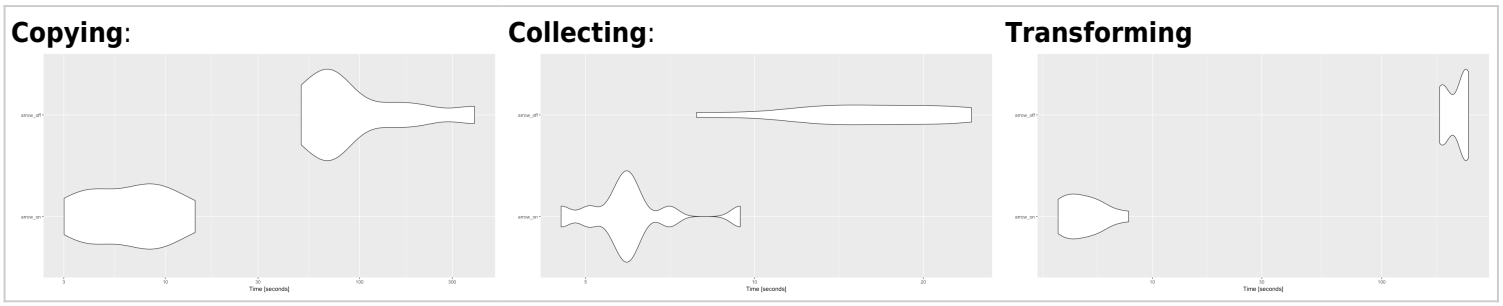


Image from here[77]

Speeding up Spark via R with Apache Arrow

Apache Arrow[78] is a cross-language development platform for in-memory data, you can read more about this in the Arrow and beyond[79] blog post. In sparklyr 1.0, we are embracing Arrow as an efficient bridge between R and Spark, conceptually.

X Axis: Time to complete task (left-hand-side is faster)
Y Axis: Top section: **WITHOUT** Apache Arrow), Bottom section: **WITH** apache Arrow.



From: https://arrow.apache.org/blog/2019/01/25/r-spark-improvements/[80]

(6) Addendum: Diskframe as potential alernative for medium sized projects?

- `disk.frame` package[81], answer to: how do I manipulate structured tabular data that doesn't fit into Random Access Memory (RAM)?.
- It makes use of two simple ideas 

    1. split up a larger-than-RAM dataset into chunks and store each chunk in a separate file inside a folder and
    2. provide a convenient API to manipulate these chunks
- It performs a similar role to distributed systems such as Apache Spark, Python's Dask, and Julia's JuliaDB.jl for medium data which are datasets that are too large for RAM but not quite large enough to qualify as *big data*.

More information: here[82] and here[83]

References

- TheRinSpark Book: https://therinspark.com/[84]
- Cheatsheets at https://www.rstudio.com/resources/cheatsheets/[85]
    - And some in Spanish Cheatsheets URL[86] > Spanish Translations – Traducciones en español
- Playground for R & Spark (for free): https://community.cloud.databricks.com/[87] Pricing[88]
- Rstudio Cloud (with free plan): https://rstudio.cloud[89]
- Webinars & scripts for hands-on practising:
    - https://rstudio.com/collections/working-with-spark/[90]
    - https://gitlab.com/radup/curs-r-introduccio/blob/master/codi/extra.tips.bigdata.R[91]
    - https://diskframe.com/articles/intro-disk-frame.html[92]

Thanks

[93]

xavier.depedro (a) seeds4c.org

Unless elsewhere noted, contents of this web site are released under a Creative Commons[94] license.

[1] https://www.slideshare.net/xavi

[2] https://blog.rstudio.com/2019/03/15/sparklyr-1-0/

[3] https://www.ub.edu

[4] https://www.ub.edu

[5] https://www.uoc.edu

[6] https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/municipal-data-office

[7] https://www.barcelona.cat/en/

[8] http://ueb.vhir.org/

[9] https://en.vhir.org

[10] http://ueb.vhir.org/

[11] https://en.vhir.org

[12] https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/municipal-data-office

[13] https://ajuntament.barcelona.cat/premsa/2018/02/13/barcelona-crea-una-nova-oficina-municipal-de-dades/

[14] https://www.bcn.cat/estadistica/angles/

[15] https://ajuntament.barcelona.cat/digital/en/noticia/a-website-has-been-launched-which-carries-out-analytic-monitoring-of-the-evolution-of-covid-19-in-the-city_956022

[16] https://dades.ajuntament.barcelona.cat/seguiment-covid19-bcn/

[17] https://publiccode.eu/

[18] https://ajuntament.barcelona.cat/digital/en/digital-transformation/technology-for-a-better-government/open-source-software

[19] http://ajuntament.barcelona.cat/digital/sites/default/files/LE_MesuradeGovern_EN_9en.pdf

[20] https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/cityos

[21] https://www.youtube.com/watch?v=OsbpZTzE5SI

[22] https://www.slideshare.net/francescabria/bria-francesca-bcn-open-source-agile-digital-transformation-strategy

[23] https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN

[24] https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN/blob/master/doc/COS1%20-%20SP12_Disseny%20de%20Seguretat%20del%20sistema%20inform%C3%A0tic.docx

[25] https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN

[26] https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN/blob/master/doc/COS1%20-%20SP12_Disseny%20de%20Seguretat%20del%20sistema%20inform%C3%A0tic.docx

[27] https://cran.r-project.org/web/packages/tidyverse/vignettes/paper.html

[28] https://cran.r-project.org/src/contrib/Archive/tidyverse/

[29] http://r4stats.com/workshops/introduction-to-modern-r/

[30] https://www.infoq.com/podcasts/martin-hadley-r-ecosystem

[31] https://www.amazon.es/Modern-Science-Chapman-Texts-Statistical-ebook/dp/B06XPNZ4H2

[32] https://b-rodrigues.github.io/modern_R/

[33] https://www.datascienceworkshops.com/catalogue/modern-r-and-the-tidyverse/

[34] https://www.youtube.com/watch?v=a-Yb518580c

[35] https://moderndive.com/

[36] https://rviews.rstudio.com/2017/06/08/what-is-the-tidyverse/

[37] https://github.com/rstudio/webinars/tree/master/55-ciencia-de-datos-R

[38] https://cran.r-project.org/web/packages/tidyverse/vignettes/paper.html

[39] https://raw.githubusercontent.com/rstudio/webinars/master/05-Data-Wrangling-with-R-and-RStudio/wrangling-

webinar.pdf
[40] https://raw.githubusercontent.com/rstudio/webinars/master/55-ciencia-de-datos-R/ciencia-de-datos-R.pdf
[41] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[42] https://github.com/garrettgman
[43] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[44] https://github.com/garrettgman
[45] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[46] https://github.com/garrettgman
[47] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[48] https://github.com/garrettgman
[49] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[50] https://github.com/garrettgman
[51] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[52] https://github.com/garrettgman
[53] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[54] https://github.com/garrettgman
[55] https://github.com/rstudio/webinars/blob/master/14-Work-with-big-data/14-Work-with-big-data.pdf
[56] https://github.com/garrettgman
[57] https://spark.apache.org/
[58] https://apache.org/
[59] https://github.com/rstudio/webinars/blob/master/42-Introduction%20to%20sparklyr/Introducing%20sparklyr%20-%20Webinar.pdf
[60] https://therinspark.com/
[61] https://therinspark.com/
[62] https://therinspark.com/
[63] https://therinspark.com/
[64] https://therinspark.com/
[65] https://svn.apache.org/repos/asf/hadoop/logos/out_rgb/
[66] https://therinspark.com/
[67] https://commons.wikimedia.org/wiki/User:Amitie_10g
[68] http://www.apache.org/licenses/LICENSE-2.0
[69] https://therinspark.com/
[70] https://therinspark.com/
[71] https://therinspark.com/
[72] https://therinspark.com/
[73] http://svn.apache.org/repos/asf/hadoop/logos/asf_hadoop/
[74] https://www.eddjberry.com/post/2017-12-05-sparkr-vs-sparklyr/
[75] https://spark.rstudio.com/
[76] https://www.slideshare.net/ICTeam/sparklyr-big-data-enabler-for-r-users
[77] https://www.slideshare.net/ICTeam/sparklyr-big-data-enabler-for-r-users
[78] https://arrow.apache.org/
[79] https://blog.rstudio.com/2018/04/19/arrow-and-beyond/
[80] https://arrow.apache.org/blog/2019/01/25/r-spark-improvements/

[81] https://github.com/xiaodaigh/disk.frame

[82] https://github.com/xiaodaigh/disk.frame

[83] https://diskframe.com/articles/intro-disk-frame.html

[84] https://therinspark.com/

[85] https://www.rstudio.com/resources/cheatsheets/

[86] https://www.rstudio.com/resources/cheatsheets/

[87] https://community.cloud.databricks.com/

[88] https://databricks.com/product/aws-pricing

[89] https://rstudio.cloud

[90] https://rstudio.com/collections/working-with-spark/

[91] https://gitlab.com/radup/curs-r-introduccio/blob/master/codi/extra.tips.bigdata.R

[92] https://diskframe.com/articles/intro-disk-frame.html

[93] http://creativecommons.org/licenses/by-sa/3.0/

[94] http://creativecommons.org/licenses/by-sa/3.0/