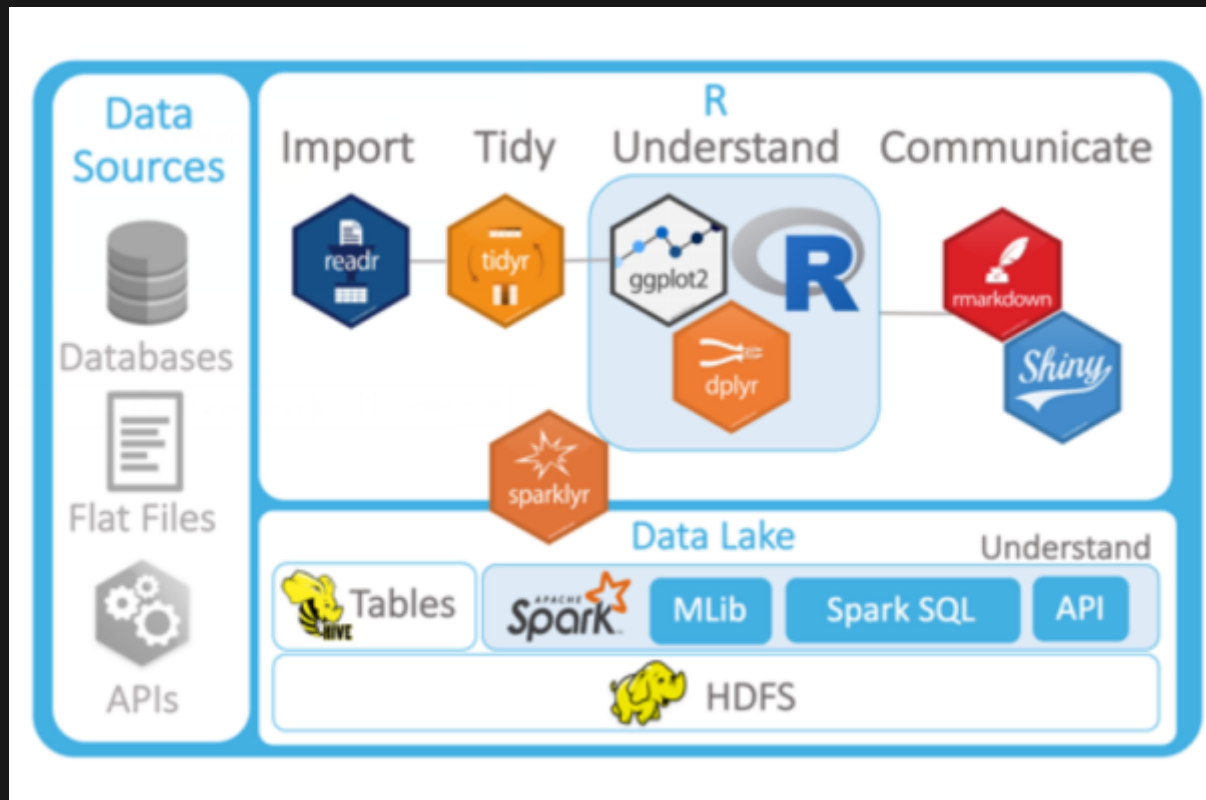


Big data with Modern R & Spark



Wednesday, June 12, 2024. 16:00-17:00h + 30' questions

Within the context of the postgraduate course on

"Data Science. Applications to Biology and Medicine with Python and R"

at University of Barcelona (IL3). 2024.

Outline

1. About me
2. Barcelona City Council Case
3. Modern R: Tidyverse
4. Modern R & Big Data
5. Spark (Apache) & Sparklyr (RStudio)
6. Sparklyr - next steps
7. Alternative approach
8. Former hands-on exercise nowadays with sparklyr

(1) About me

Xavier de Pedro Puente, Ph.D. [xavier.depedro \(a\) seeds4c.org](mailto:xavier.depedro@seeds4c.org)

- **Academics:**

- Degree in Biology (UB^[1])
- Ph.D. in Ecology (UB^[2])
- Postgraduate in Bioinformatics (UOC^[3])

- **Current Work:**

- Senior technician at **Climate Change and Sustainability Office** (Barcelona City Council^[4])

- **Past (related) Work:**

- Senior technician at Development Division, (IMI^[5], Barcelona City Council^[6])
- Senior technician at Municipal Data Office (OMD^[7], Barcelona City Council^[8])
- Bioinformatics technician (UEB^[9], VHIR^[10])
- Systems administrator (UEB^[11], VHIR^[12])



(2) Barcelona City Council Case

- 2018: OMD - Municipal Data Office opened^[13] (details^[14]):
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^[15] portal
- Recent product^[16]: Covid-19 in BCN monitoring with an R Shiny App^[17]

Indicador

COVID-19

Daily evolution **new**

Cumulative evolution

Territorial distribution

Deaths **new**

Weather and Environmental Quality

Estimated population

Mobility and transports

Port and airport **new**

Public transportation and vehicles

Visitors in Barcelona

Daily commuting

Economy

Prices

Labour market

Consumption and Stocks market

Housing

Tourism **new**

Opinion

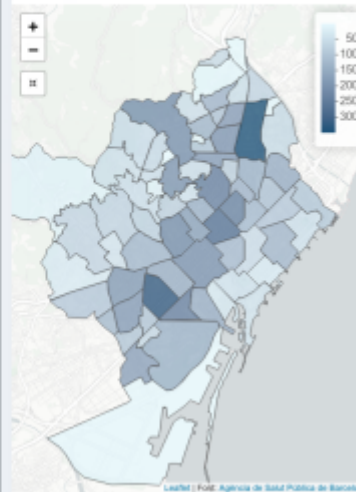
Networks

Timeline

About this site

Neighbourhood cases: number of positive cases by neighbourhood

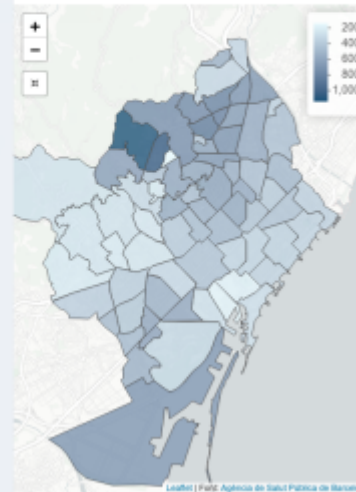
Descarrega les dades en csv



1

Incidence rate by neighbourhood: rates per 100,000 inhabitants

Descarrega les dades en csv



1

Barcelona City Council: Public money - public code

- Barcelona: first city to join the Free Software Foundation, Public Money, Public Code^[18] campaign
- One of the case studies^[19] of the use of open-source software and open code to democratise cities.

There is a Government policy for open source and agile methodologies^[20]

Open & Agile Digital Transformation Toolkit

Open digitalisation programme: free software and agile implementation of Barcelona City Council services that defines the process of profound, progressive change in the way the city will offer its services to the general public in coming years.

A decided change that puts the general public at the centre of designing "digital by default" government services focused on technological sovereignty, free software and open code, and the ethical use of data, in order to be more open and transparent, using agile methodologies and open technologies in order to be more effective, and transforming public contracting in order to promote innovation in the local labour industry, strengthening the country's small and medium-sized companies. A change that eventually leads to a more open, simple and enriching relationship with the general public as a whole.

[Download the government measure](#)

Digital services standards

A set of criteria for the development and management of digital services for citizens, in an agile, open and user-centred way.

[Download PDF](#)

Technology code of practice

A set of policies and directives mandating the use of open standards, open source and ethical data management.

[Download PDF](#)

Agile methods

Principles and practices for the agile implementation of Barcelona City Council in the delivery of digital services.

[Download PDF](#)

Technological sovereignty

The implementation of technological sovereignty in the delivery of BCB City Council's digital services.

[Download PDF](#)

ICT Procurement Handbook

The definition of a new model of relation and partnership with technology suppliers, based on the Code of Technological Practices.

[Download PDF](#)

Innovative Public Procurement

The transformation of procurement to obtain solutions that are tailored to the real needs of citizens, while helping companies to be more innovative.

[Download PDF](#)

Barcelona City Council: CityOS

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

City OS: internal data management, known as "Data Lake"^[21]

Video^[22]

OPEN SOURCE URBAN PLATFORM

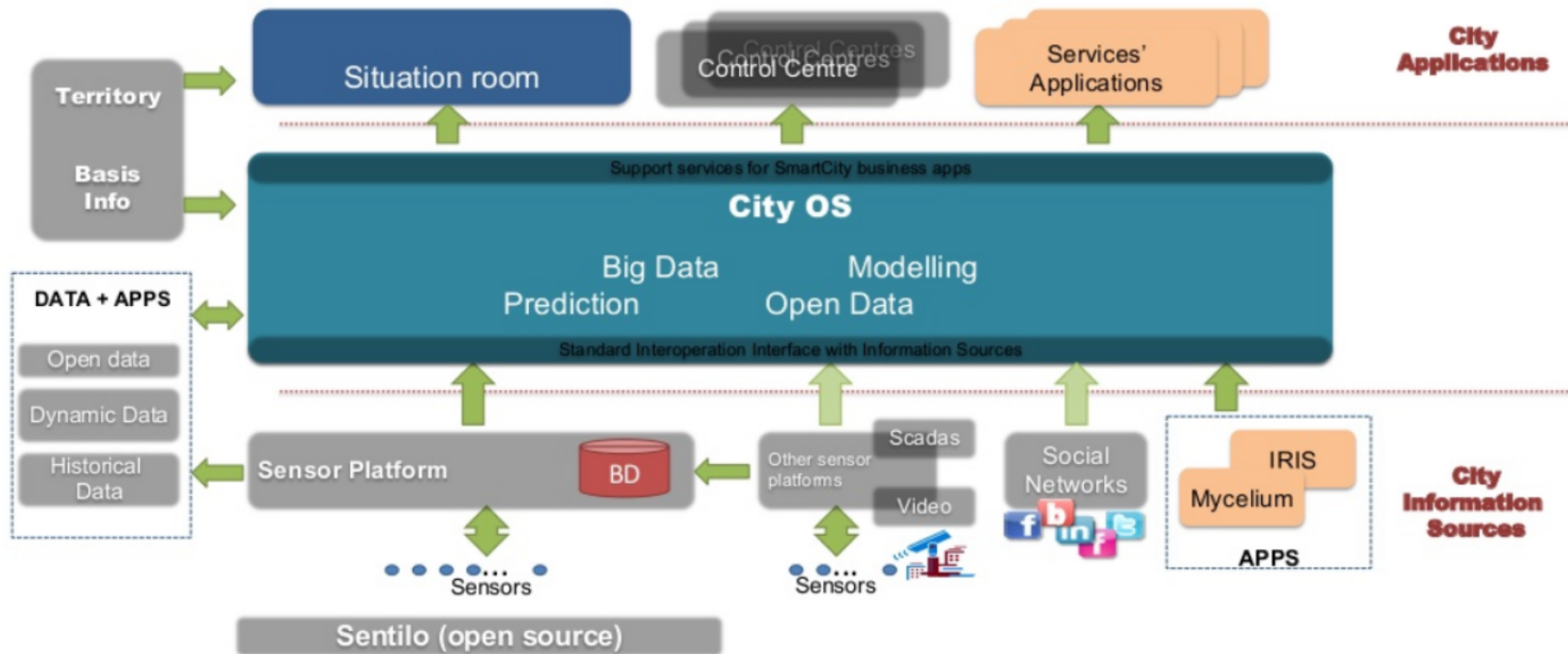
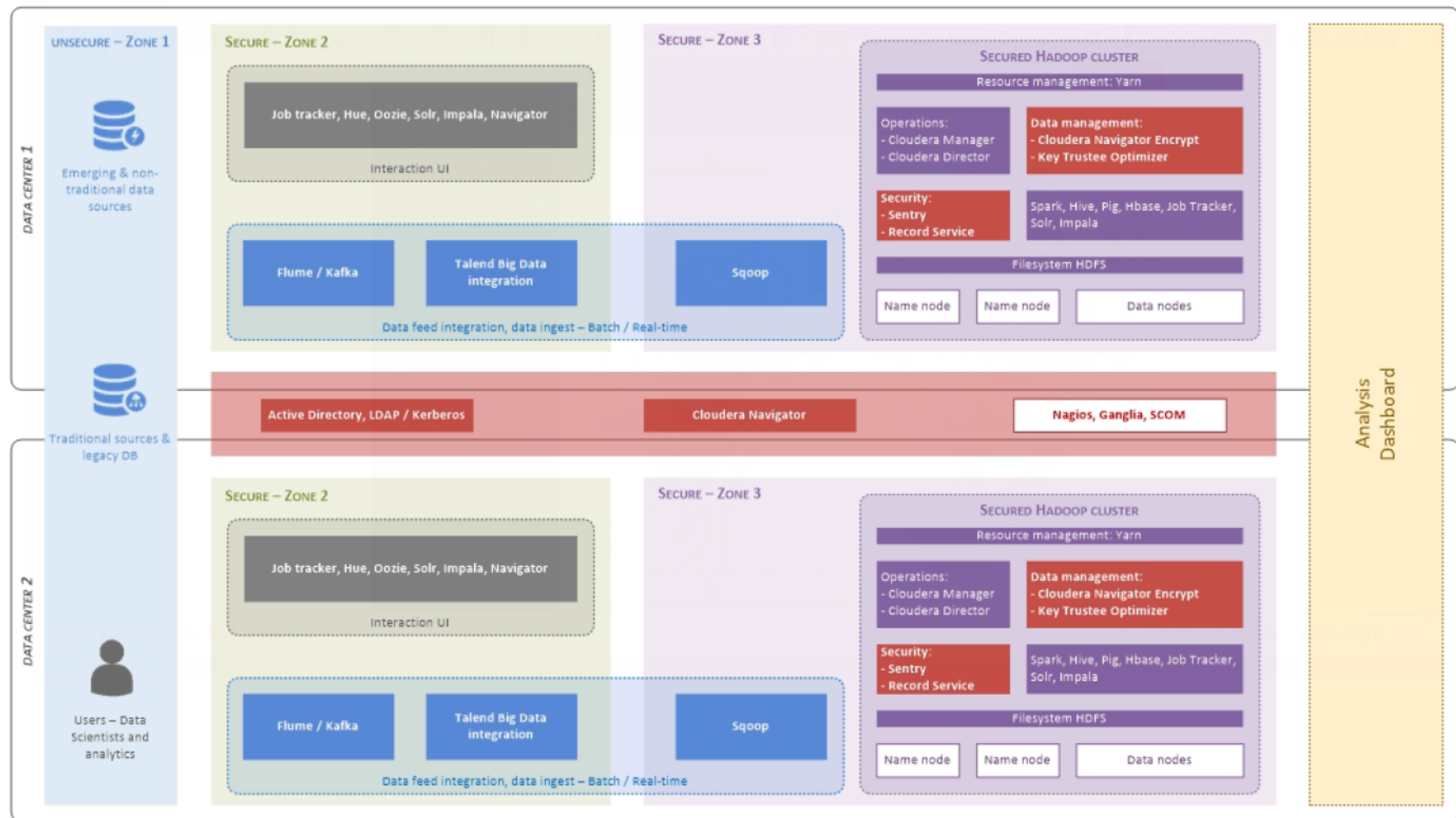


Image from Francesca Bria^[23] (Barcelona Digital City Roadmap 2017-2020)

Barcelona City Council: CityOS Technologies

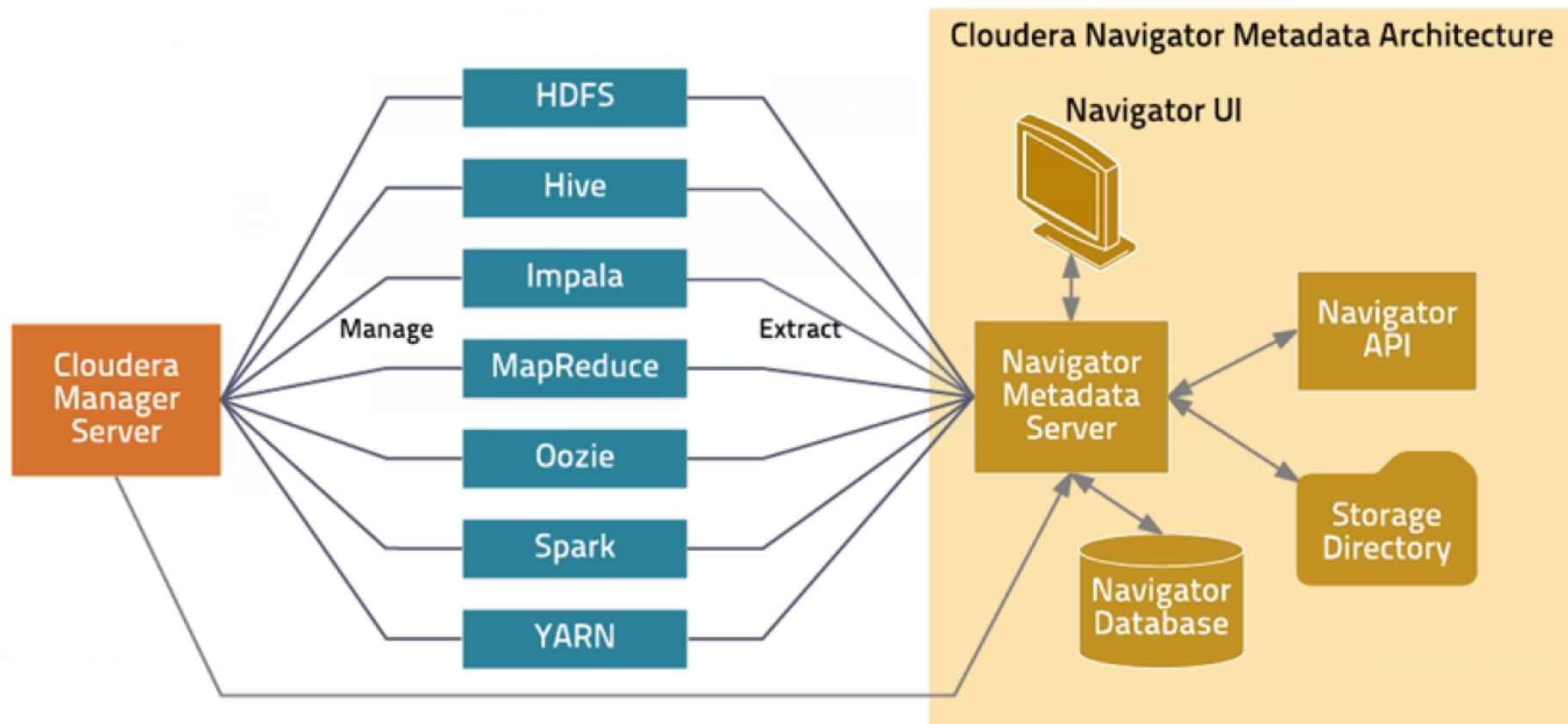
CityOS technologies: GNU/Linux, CentOS, Cloudera, Activiti, Talend, Protégé, **R**, Zabbix, Nagios, Ganglia



Source: https://github.com/AjuntamentdeBarcelona/CityOS_AjBCN^[24] - Image from J. Berdonces a Github^[25]

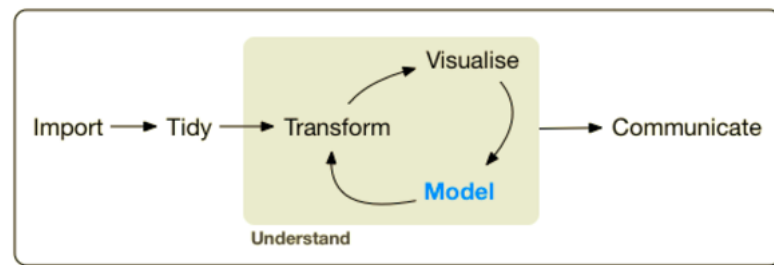
Barcelona City Council: CityOS - Cloudera Manager

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



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

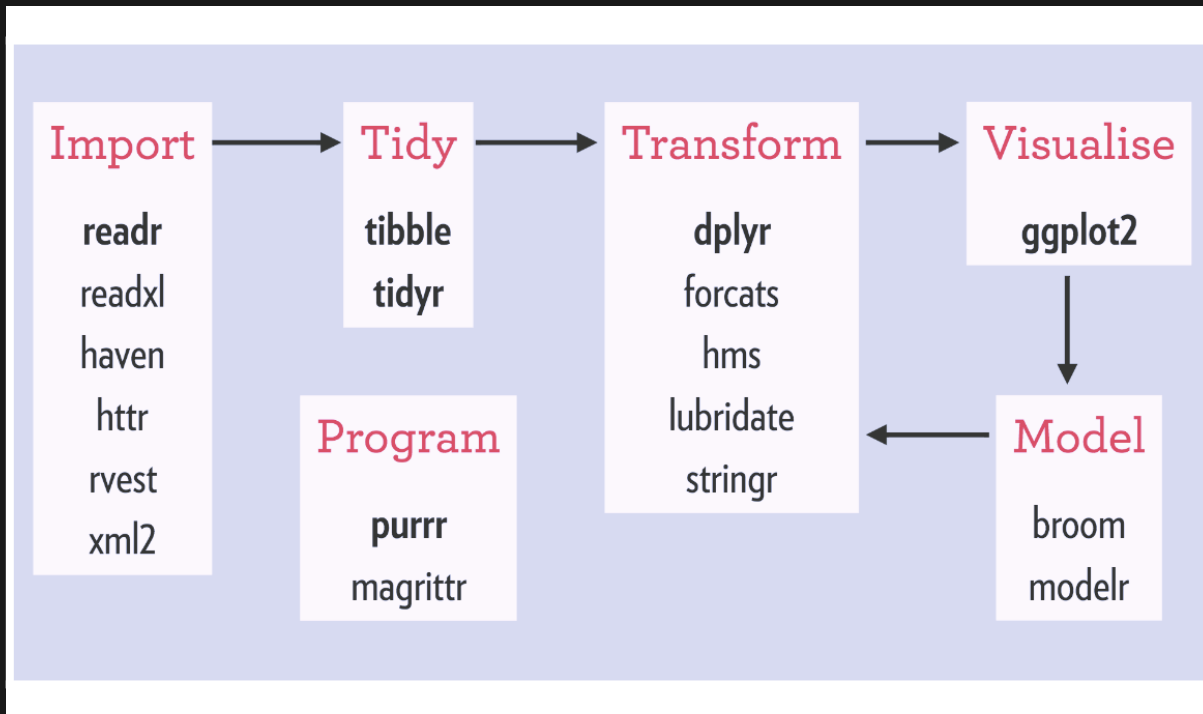
- 2000: First stable beta version (v1.0) released.
- ... "Base R" ...
- 2016: Tidyverse^[28] 1.0.0 package released on CRAN^[29]
- 2016: "Introduction to Modern R - r4stats.com"^[30]
- 2017: "Martin Hadley on R and the Modern R Ecosystem - InfoQ Podcasts"^[31]
- 2017: "Modern Data Science with R - Chapman & Hall/CRC"^[32]
- 2019: "Modern R with the tidyverse - Econometrics and Free Software"^[33] (blog)
- 2019: "Modern R and the Tidyverse - Data Science Workshops"^[34]
- 2020: "Modern R: Welcome to the tidyverse"^[35] (video tutorial)
- 2020: Statistical Inference via Data Science: ModernDive into R & Tidyverse^[36]
- 2022: Modern R with the Tidyverse^[37]



Program

Modern R (Tidyverse)

Workflow, Packages, People/Community



L'equip "tidyverse"



@hadleywickham



@jimhester_



@JennyBryan



@_lionelhenry



@krlmlr



@dataandme



@romain_francois



@drob



@thomasp85



@GaborCsardi



@LucyStats



@topepos



tidyverse.org

Derived from [here](#)^[38] & [here](#)^[39]

Modern R (Tidyverse) principles

1. Main structures are ordered data

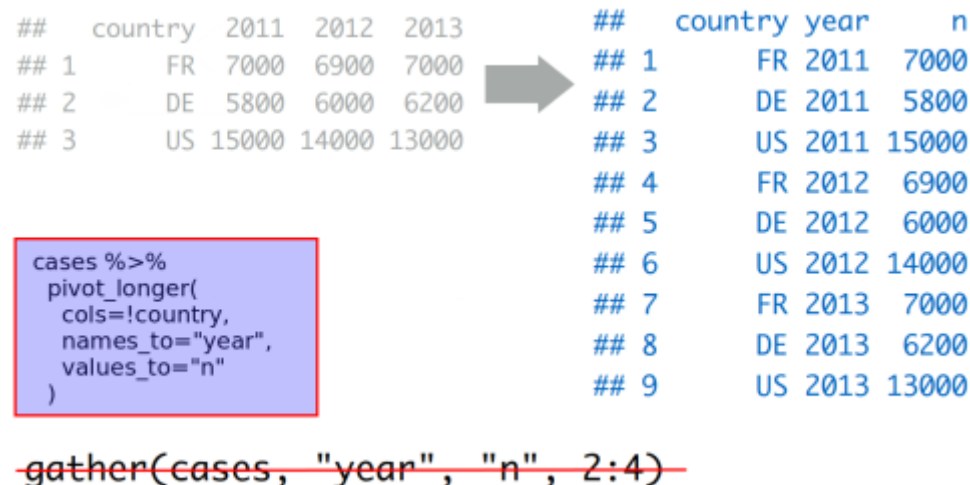
- Each variable is saved in its own column.
- Each observation is saved in its own row.
- Each "type" of observation stored in a single table

2. Each function represents one step

3. Functions are combined with the pipe operator %>%

- or with the earlier pipe^[40] from base R: |>

4. Each step is a query or command



The diagram illustrates the transformation of a wide data format into a long data format using the tidyverse pipe operator (%>%). On the left, a wide table with columns 'country', '2011', '2012', and '2013' is shown. An arrow points to the right, where a long table with columns 'country', 'year', and 'n' is displayed. A red box highlights the tidyverse code used for this transformation: `cases %>% pivot_longer(cols=!country, names_to="year", values_to="n")`. Below this, the base R code `gather(cases, "year", "n", 2:4)` is shown with a red line through it, indicating it is the older method.

```
##   country 2011 2012 2013
## 1      FR 7000 6900 7000
## 2      DE 5800 6000 6200
## 3      US 15000 14000 13000
```

→

```
##   country year    n
## 1      FR 2011  7000
## 2      DE 2011  5800
## 3      US 2011 15000
## 4      FR 2012  6900
## 5      DE 2012  6000
## 6      US 2012 14000
## 7      FR 2013  7000
## 8      DE 2013  6200
## 9      US 2013 13000
```

```
cases %>%
  pivot_longer(
    cols=!country,
    names_to="year",
    values_to="n"
  )
```

~~gather(cases, "year", "n", 2:4)~~

(4) Modern R & Big Data

Data > RAM = Problem

✈ Airlines Data Set

Arrival and departure details for all commercial flights in US between October 1987 and April 2008.

120,000,000 records. **12 GB**

stat-computing.org/dataexpo/2009/



Data does not
fit in memory

3 classes of Big Data Problems

Class 1. Extract Data

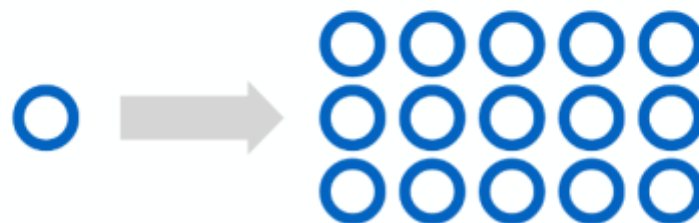
Problems that require you to extract a subset, sample, or summary from a Big Data source. You may do further analytics on the subset, and the subset might itself be quite large.



© 2015 H2Studio, Inc. All rights reserved.

Class 2. Compute on the parts

Problems that require you to repeat computation for many subgroups of the data, e.g. you need to fit one model per individual for thousands of individuals. You may combine the results once finished.



© 2015 RStudio, Inc. All rights reserved.

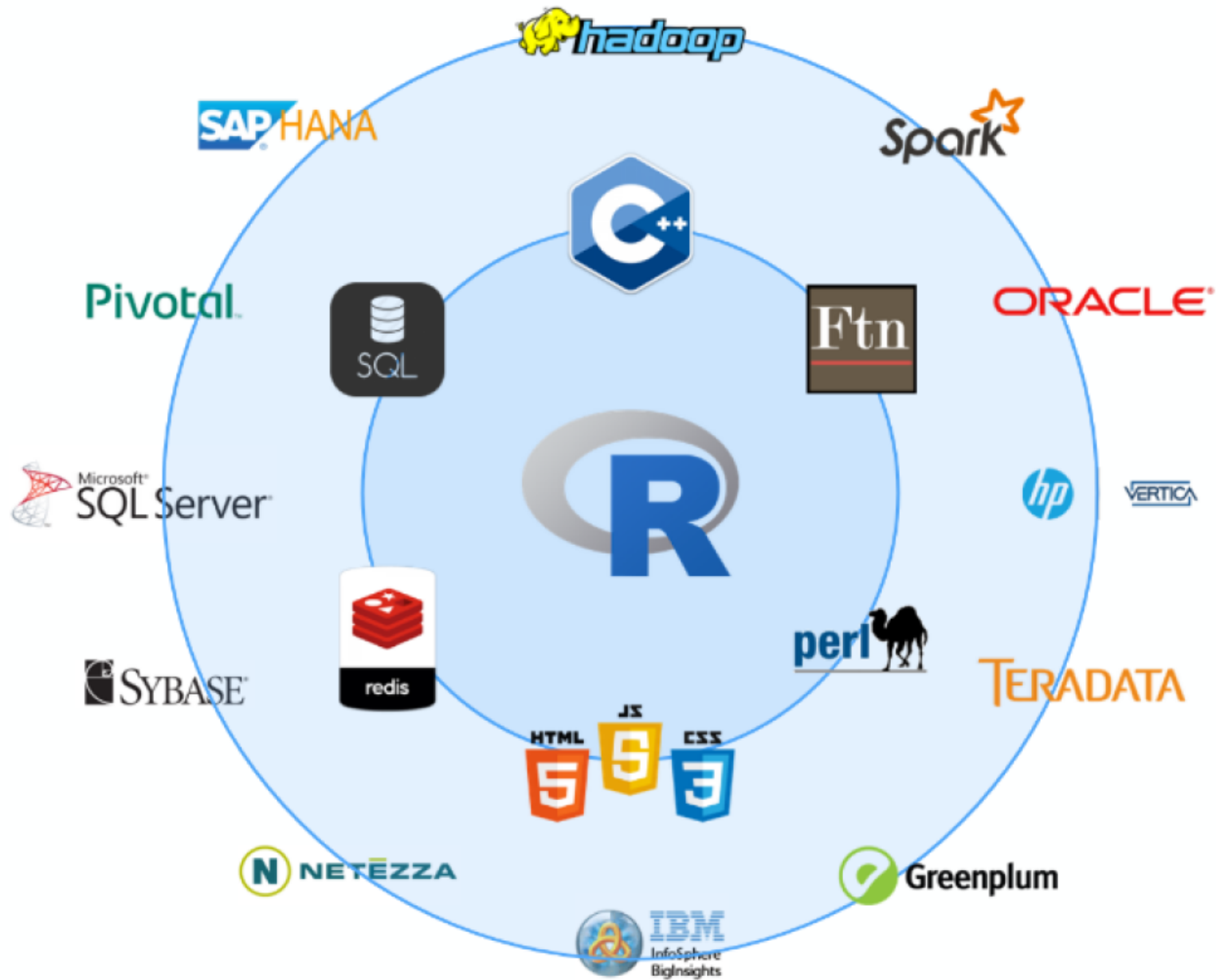
Class 3. Compute on the whole

Problems that require you to use all of the data at once. These problems are irretrievably big; they must be run at scale within the data warehouse.



© 2015 RStudio, Inc. All rights reserved.

R interacts with other technologies



© 2015 RStudio, Inc. All rights reserved.

R General Strategy (i): Class 1 & 2

General Strategy

Store big data in a data warehouse

1. Pass subsets of data from warehouse to R
2. Transform R code, pass to warehouse.



© 2015 RStudio, Inc. All rights reserved.

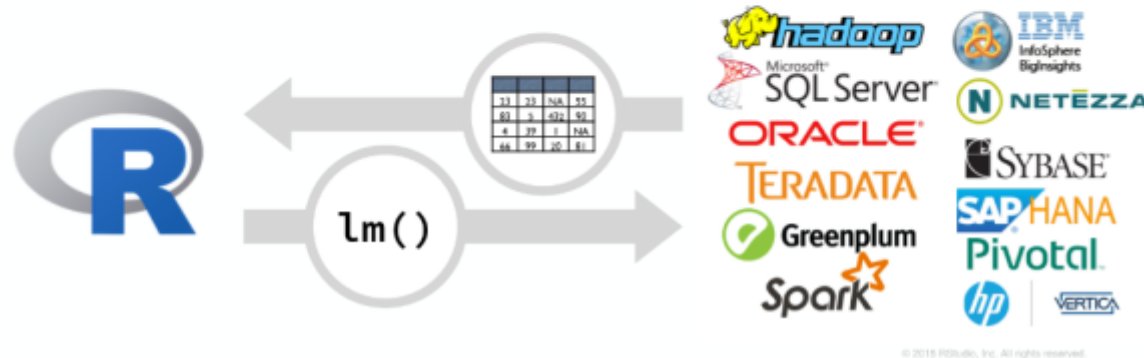
From an RStudio seminar^[50] by Garrett Grolemund^[51]

R General Strategy (ii): Class 1 & 2

General Strategy

Store big data in a data warehouse

1. Pass subsets of data from warehouse to R
2. Transform R code, pass to warehouse.



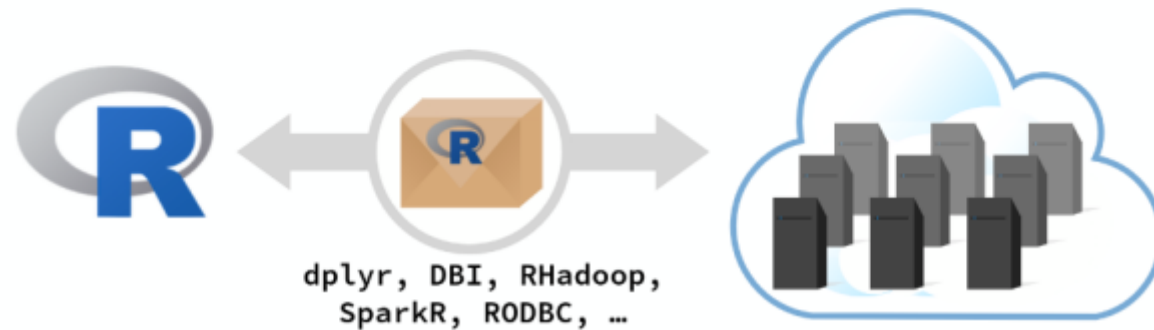
From an RStudio seminar^[52] by Garrett Grolemund^[53]

R General Strategy (iii): Class 3

General Strategy

Store big data in a data warehouse

1. Pass subsets of data from warehouse to R
2. Transform R code, pass to warehouse.



© 2015 PDS, Inc. All rights reserved.

From an RStudio seminar^[54] by Garrett Grolemond^[55]

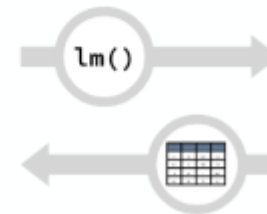
Class 3 - dbplyr

R - tidyverse - dplyr + dbplyr



Package that provides data manipulation syntax for R. Comes with **built-in SQL backend**:

1. **Connects** to DBMS's
2. **Transforms R code** to SQL, sends to DBMS to run in DBMS
3. **Collect results** into R



© 2015 RStudio, Inc. All rights reserved.

R General Strategy (iv): Recap

Recap: Access Big Data



Store data outside of memory in data warehouse



Use an R package as an API to the data warehouse. dplyr, DBI, sparkR, others.

```
db <-
```

Create and work with connection object

```
rm(db)  
gc()
```

Close connection when finished

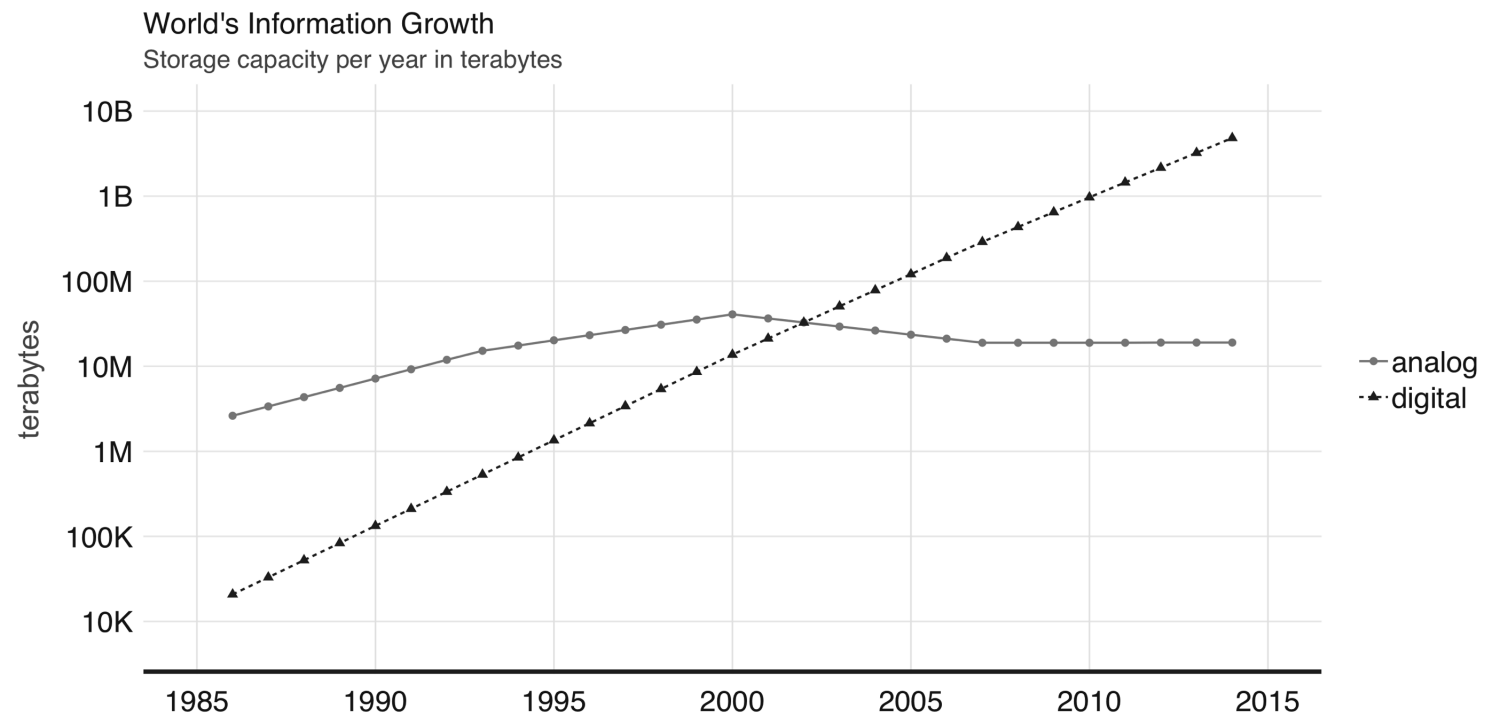
© 2015 RStudio, Inc. All rights reserved.

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

- Spark^[60], from Apache Foundation^[61]: 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?**

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)^[64]

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)^[65]

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)^[66]

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.



Hive | Facebook - 2008

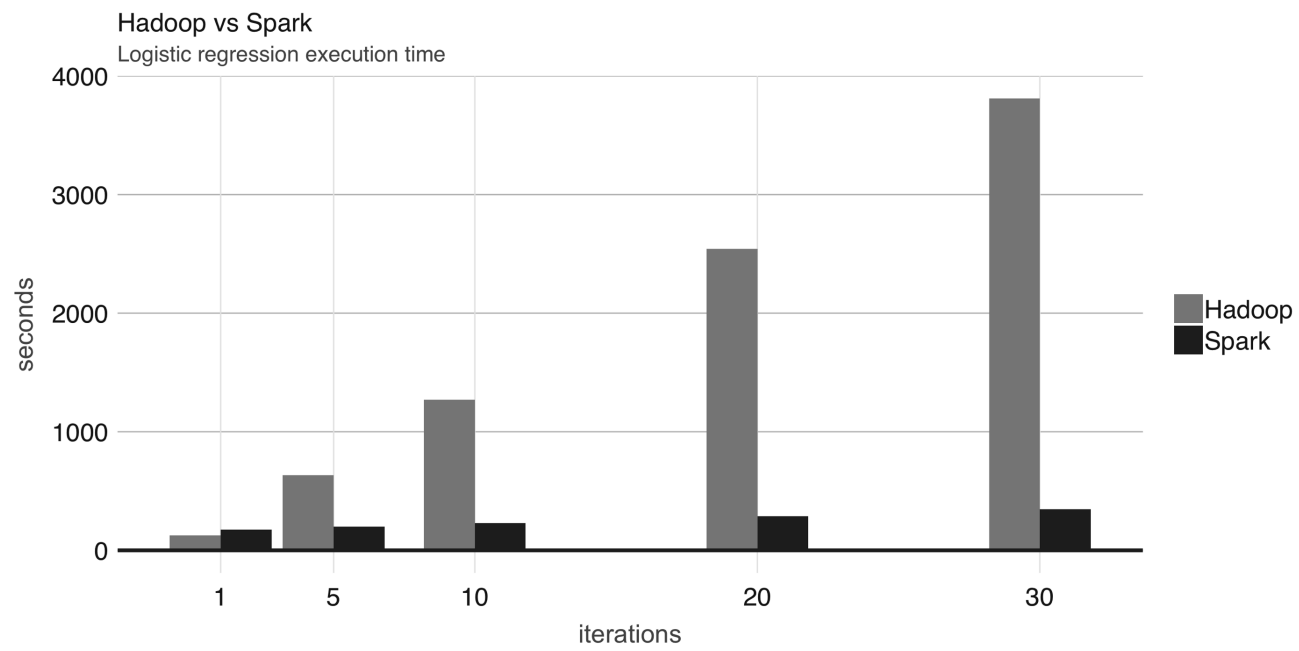
- Hadoop provided support to perform MapReduce operations over a distributed file system, but it still required MapReduce operations to be written with code every time a data analysis was run
- **Hive** project (2008, by Facebook) brought **Structured Query Language** (SQL) support to Hadoop.
 - Data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



From "The R In Spark" (book)^[69] | Image from Davod^[70] with Apache License 2.0^[71]

Spark (closed sourced) | UCBerkeley - 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)^[72]

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)^[73]

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.

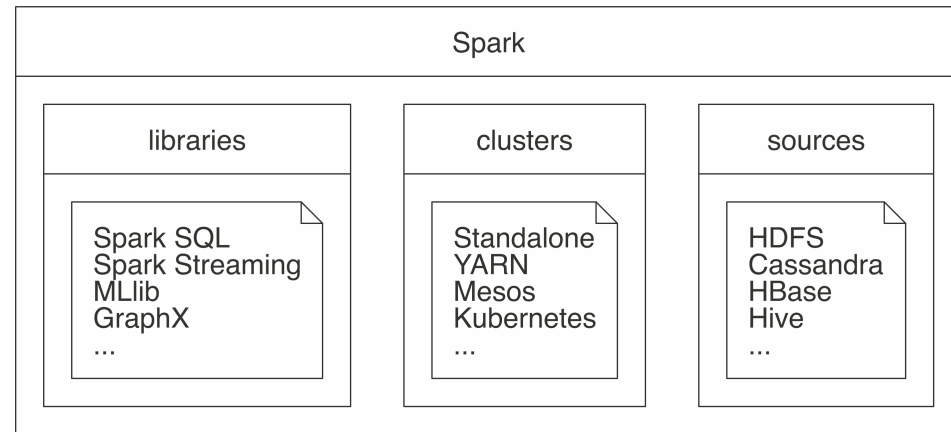
	Hadoop Record	Spark Record
Data Size	102.5 TB	100 TB
Elapsed Time	72 mins	23 mins

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)^[74]

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)^[75] | Image from the Apache Software Foundation, with Apache License 2.0^[76]

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

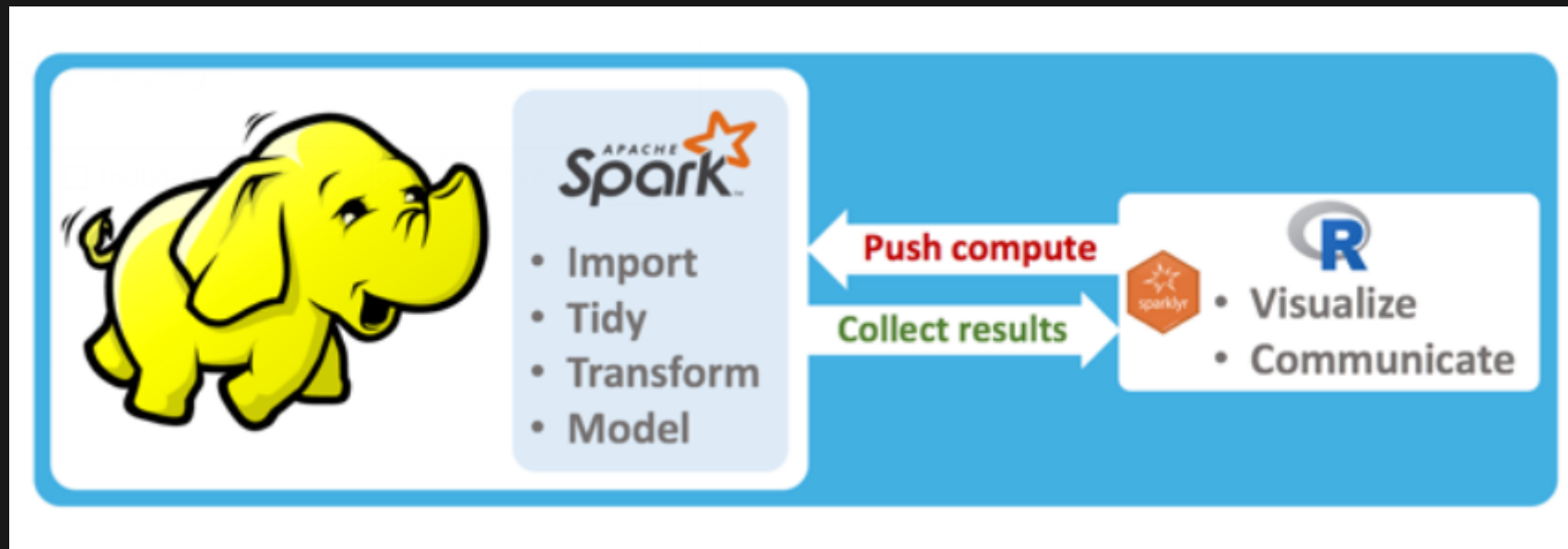
Feature	SparkR	sparklyr
Data input & output	+ +	+ +
Data manipulation	-	+ + +
Documentation	+ +	+ +
Ease of setup	+ +	+ +
Function naming	- -	+ + +

Installation	+	+	+
Machine learning	+	+	+
Range of functions	+	+	+
Running arbitrary code	+	+	+
Tidyverse compatability	-	-	-

From <https://www.eddjberry.com/post/2017-12-05-sparkr-vs-sparklyr/>^[77]

Sparklyr | RStudio

- Sparklyr, from Rstudio: <https://spark.rstudio.com/>^[78]
- 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



Summary: Big Data with Modern R & Spark

Big Data with Modern R & Spark in context

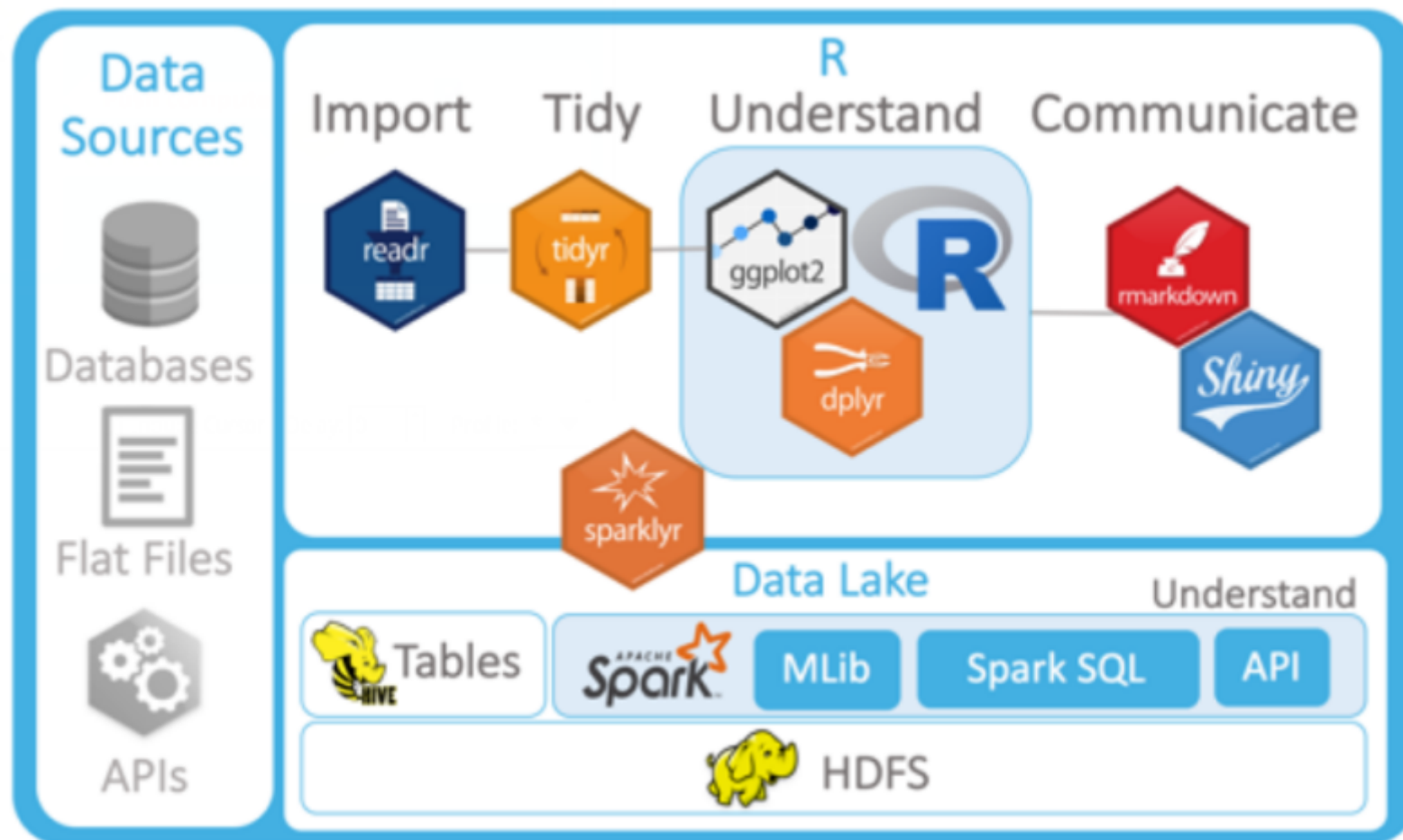


Image from here^[80]

Speeding up Spark via R with Apache Arrow

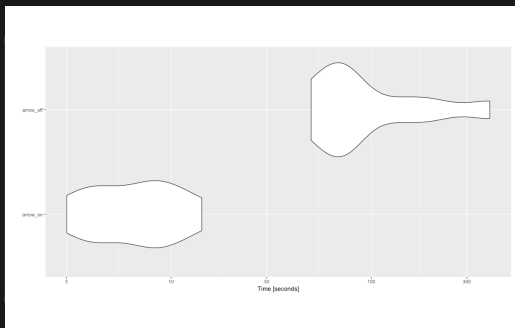
Apache Arrow^[81] is a cross-language development platform for in-memory data, you can read more about this in the Arrow and beyond^[82] blog post. In sparklyr 1.0, we are embracing Arrow as an efficient bridge between R and Spark, conceptually.



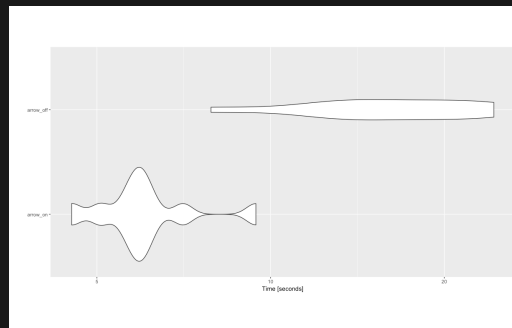
X Axis: Time to complete task (left-hand-side is faster)

Y Axis: Top section: **WITHOUT** Apache Arrow), Bottom section: **WITH** apache Arrow.

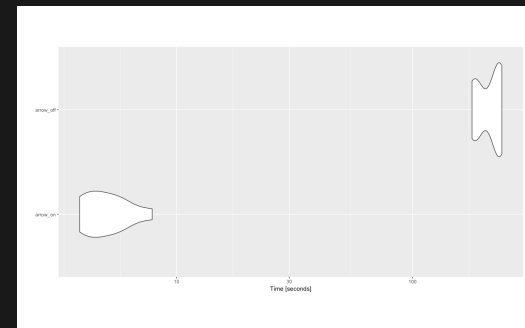
Copying:



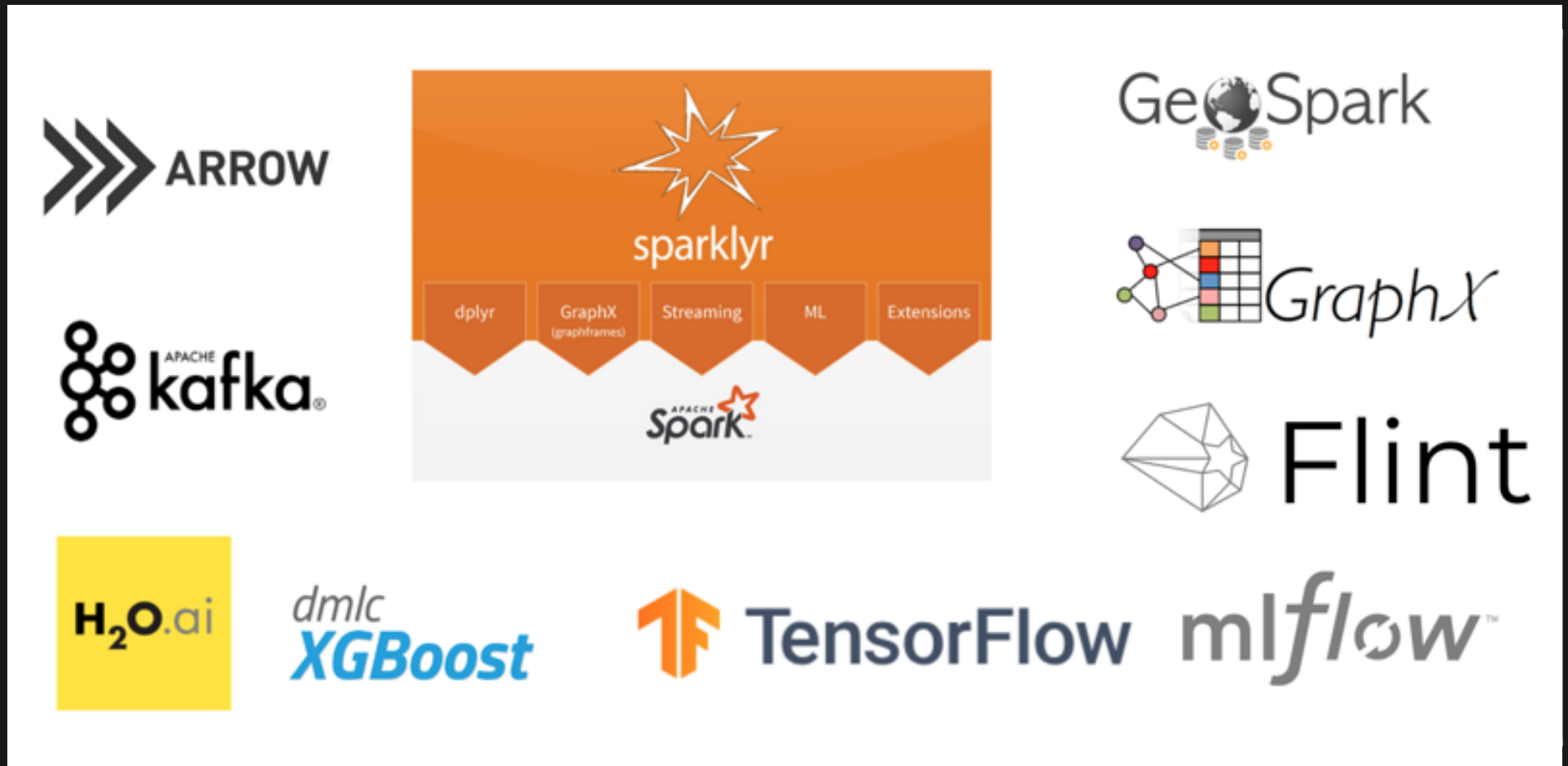
Collecting:



Transforming



(6) Sparklyr - next steps



From: <https://github.com/harryprince/awesome-sparklyr>^[84]

(7) Addendum: Diskframe as potential alternative for medium sized projects?

- `disk.frame` package^[85], 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^{\[86\]}](#) and [here^{\[87\]}](#)

(8) Former hands-on exercise nowadays with sparklyr

Remember the hands on exercise we did on the session about "**Reproducible Work in Data Science**" in this postgraduate course?

We will revisit the work we did in that session, and we will evolve the Rmd you did in order to adapt it to use sparklyr to achieve the goal we had in that hands-on exercise.

We can't make use of the posit.cloud free plan since the Spark infrastructure we need does require more than just 1 Gb of RAM (the RAM provided for free in posit.cloud free plan).

So that in this seminar, we will see the job performed in datascience.seeds4c.org, which is a lxc (linux container) based on Ubuntu 20.04 LTS with 4Gb of RAM and 4 cpu.

<http://datascience.seeds4c.org:8787/>^[88]

117 Mb (and 63 secs)	sparklyr	1560 Mb
(and 165 secs)	just R (not involving Spark)	

References

- TheRinSpark Book: <https://therinspark.com/>^[90]
- Cheatsheets at <https://www.rstudio.com/resources/cheatsheets/>^[91]
 - And some in Spanish Cheatsheets URL^[92] > Spanish Translations – Traducciones en español
- Posit (Rstudio) Cloud (with free plan): <https://posit.cloud/plans>^[93]
- Webinars & scripts for hands-on practising:
 - <https://spark.rstudio.com/get-started/>^[94]
 - <https://gitlab.com/radup/curs-r-introduccio/blob/master/codi/extra.tips.bigdata.R>^[95]
 - <https://diskframe.com/articles/02-intro-disk-frame.html>^[96]
- Compilation of links related to sparklyr:
<https://github.com/harryprince/awesome-sparklyr>^[97]
- Playground for R & Spark (14d for free):
<https://community.cloud.databricks.com/>^[98] Pricing^[99]
- Tidyverse Skills for Data Science. Carrie Wright, Shannon E. Ellis, Stephanie C. Hicks and Roger D. Peng. 2021-09-02

- <https://jhudatascience.org/tidyversecourse/>^[100]

Thanks

[101]

[xavier.depedro \(a\) seeds4c.org](mailto:xavier.depedro@seeds4c.org)

Unless elsewhere noted, contents of this web site are released under a Creative Commons^[102] license.

Earlier videorecording

[+]

<https://www.ub.edu>

<https://www.ub.edu>

<https://www.uoc.edu>

<https://www.barcelona.cat/en/>

<https://ajuntament.barcelona.cat/imi/en>

<https://www.barcelona.cat/en/>

<https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/municipal-data-office>

<https://www.barcelona.cat/en/>

<http://ueb.vhir.org/>

<https://en.vhir.org>

<http://ueb.vhir.org/>

<https://en.vhir.org>

<https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/municipal-data-office>

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

<https://www.bcn.cat/estadistica/angles/>

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

<https://dades.ajuntament.barcelona.cat/seguiment-covid19-bcn/>

<https://publiccode.eu/>

<https://ajuntament.barcelona.cat/digital/en/digital-transformation/technology-for-a-better-government/open-source-software>

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

<https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/cityos>

<https://www.youtube.com/watch?v=OsbpZTzE5SI>

