Rusers Meetup - Una Interface a la Ciencia de Datos

Pontificia Universidad Javeriana

Calle 18 No 118-250 Cali, Valle del Cauca

(Edificio las Palmas - Sala de Proyección # 1)

Sábado 19 de Agosto de 2017

8:30 - 13:00

- *Apertura 8:30 am - 8:42am*
- *Open Data + Mongo DB + R*
  * *(Jhon Carvajal, Victor Males) 8:45 - 9:45am*
- *Tips Rápidos para visualizaciones de Calidad con ggplot2*
  * *(Maria Isabel Arce) 9:45 - 10:45am*

Break 10:45 - 11:00 am

- *Introducción a los datos espaciales en R*
  * *(Frank Hurtado, Dave Montero) 11:00 - 12:00pm*
- *Invitación Campus Nova 12:00pm - 12:15pm*
- *Introducción al Big Data - sparklyr - Google Bigquery & Ejemplos de Machine Learning con R y Tensor Flow en la nube.*
  * *(Camilo Herrera) 12:15 - 01:00 PM*
- *Cierre 01:00 PM*

42Data Turn data into smart information

COMITÉ DE SOFTWARE LIBRE DE CALI

Pontificia Universidad JAVERIANA Cali

Introducción al Big Data & Ejemplos de Machine Learning con R

Camilo A Herrera R

Estadístico - Universidad del Valle

Sp. Data Science & Sp. Executive Data Science - Johns Hopkins University

Msc candidate - Biometría - Universidad de Buenos Aires

twitter: @hr_mr_zork - web: http://camiloherrera.co/

Email: ch@camiloherrera.co

Define: Big Data.

Velocity

Volume

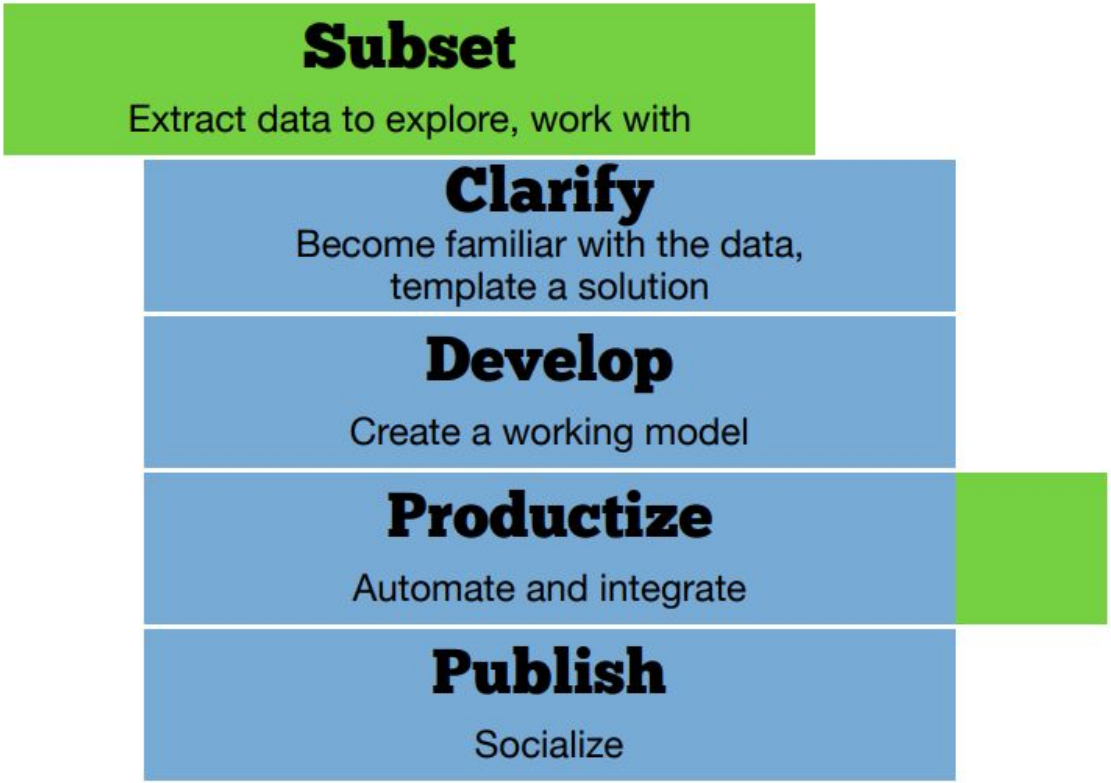
Value

Variety

Veracity

Data > RAM

Ciclo de Vida de un Proyecto de Analitica

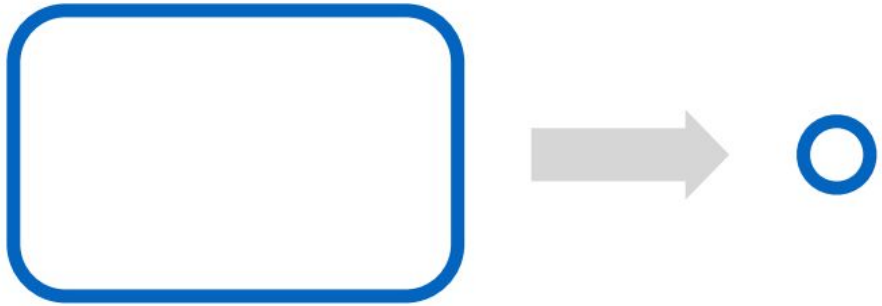


¿Cómo utilizan los analistas el Big Data?

Problemas Analíticos con Big Data

Clase1: Extraer Datos:

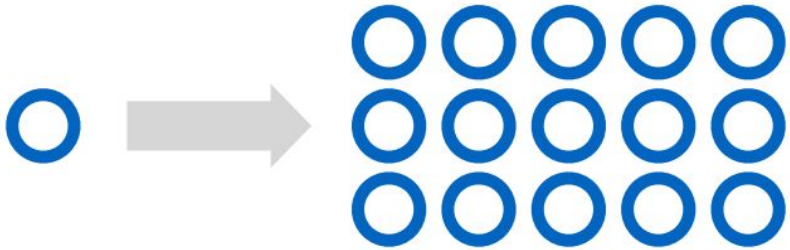
Problemas que requieren extraer un subconjunto, una muestra o un resumen de una fuente de datos grande. Puede hacer más análisis en el subconjunto, y el subconjunto podría ser bastante grande.



Problemas Analíticos con Big Data

Clase 2: Calcular en Partes:

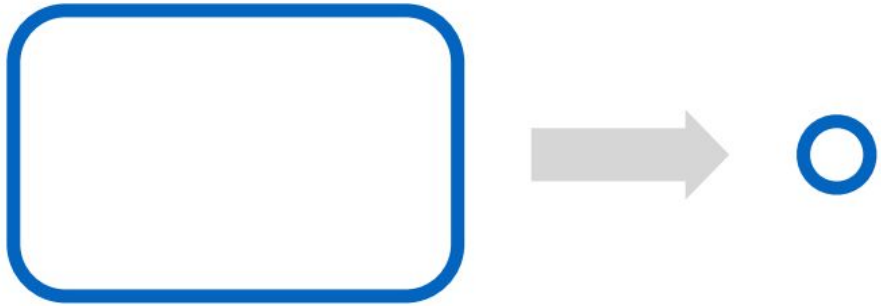
Problemas que requieren que se repita el cálculo para muchos subgrupos de los datos, Por Ejemplo si usted necesita encajar un modelo por individuo para los millares de individuos. Usted puede combinar los resultados una vez terminado.



Problemas Analíticos con Big Data

Clase 3: Calcular en Conjunto:

Problemas que requieren que utilice todos los datos a la vez. Estos problemas son irremediablemente grandes; Se deben ejecutar a escala dentro de un almacén de datos.



Ciclo de Vida de un Proyecto de Analitica

Subset

Extract data to explore, work with

Clarify

Become familiar with the data, template a solution

Develop

Create a working model

Productize

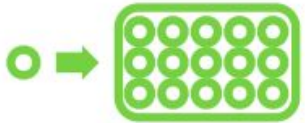
Automate and integrate

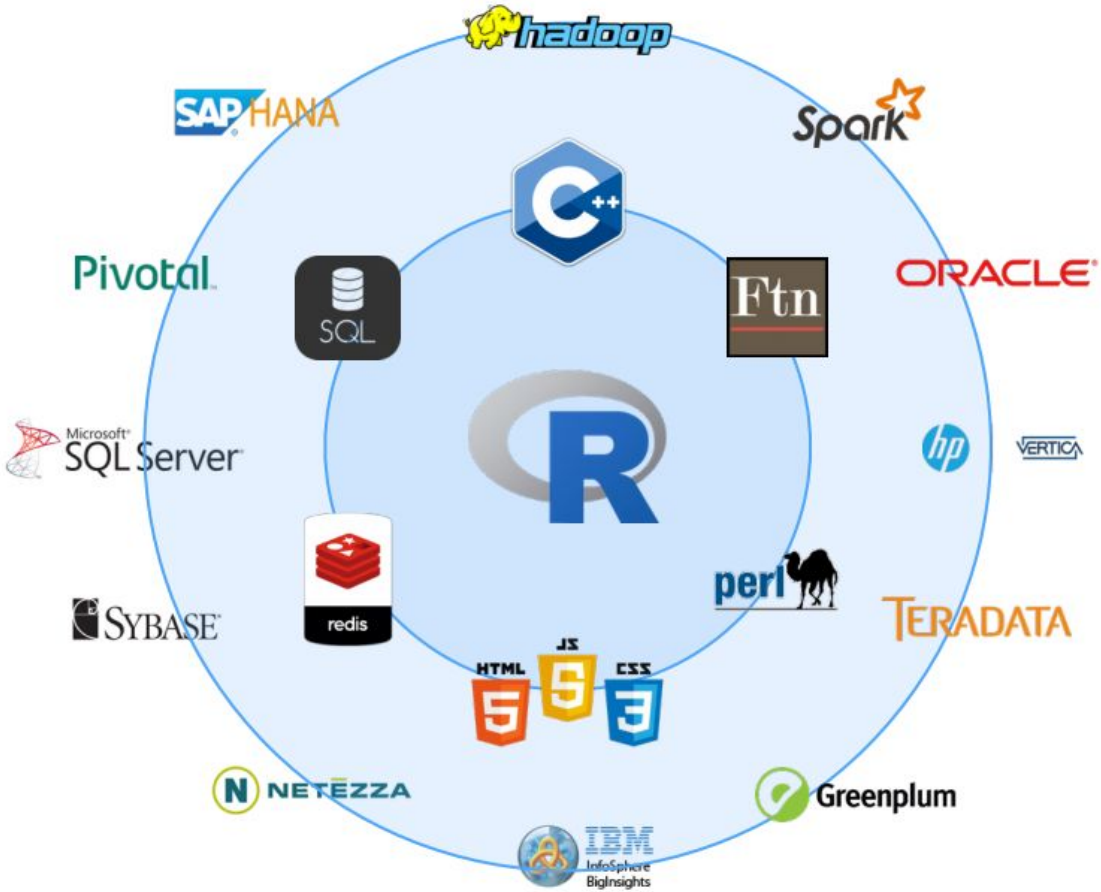
Publish

Socialize

Class 1

Class 2
Class 3

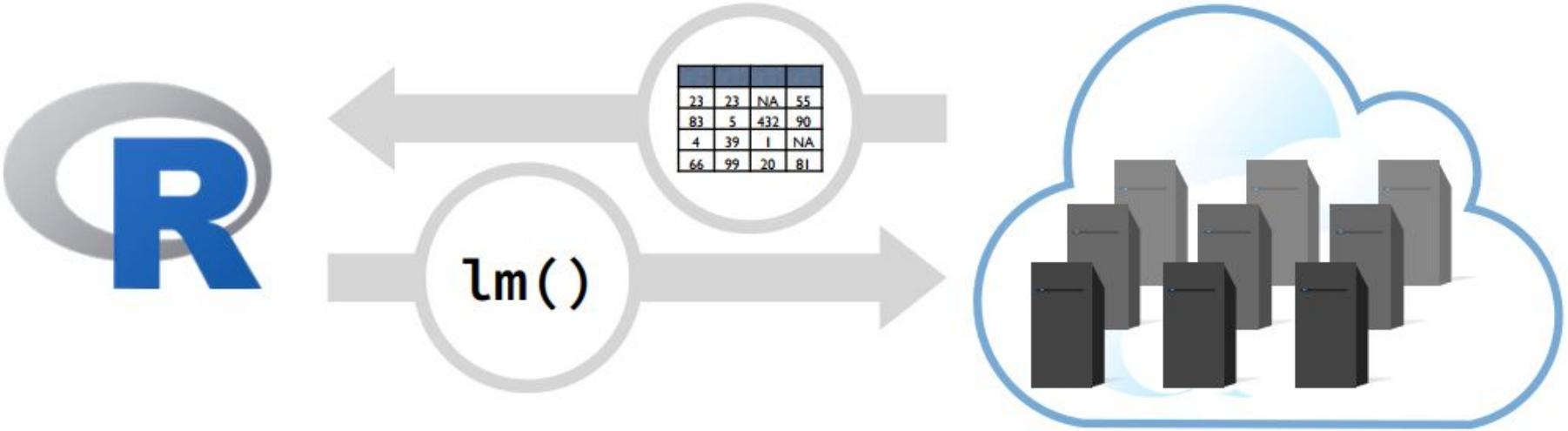



hadoop
SAP HANA
Spark
C++
Pivotal
SQL
Ftn
ORACLE
Microsoft SQL Server
R
hp
VERTICA
redis
perl
SYBASE
TERADATA
HTML
JS
CSS
NETEZZA
Greenplum
IBM
InfoSphere
BigInsights

Estrategia General:

Guardar datos en un data warehouse

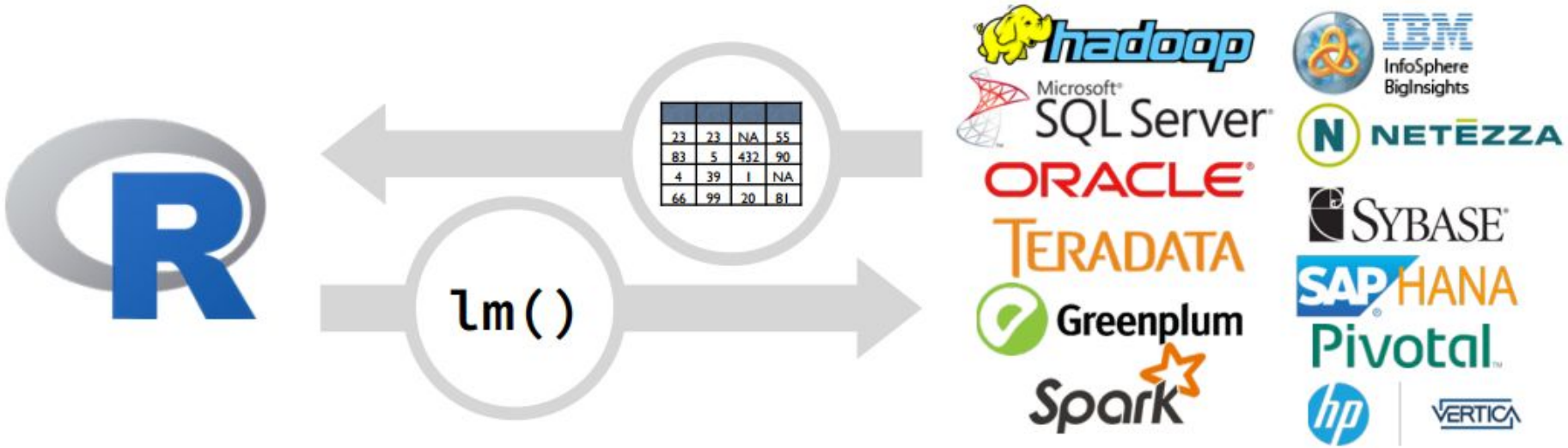
1. Pasar los subconjuntos de datos del warehouse a R.
2. Transformar el código R, pasarlo al warehouse.



Estrategia General:

Guardar datos en un data warehouse

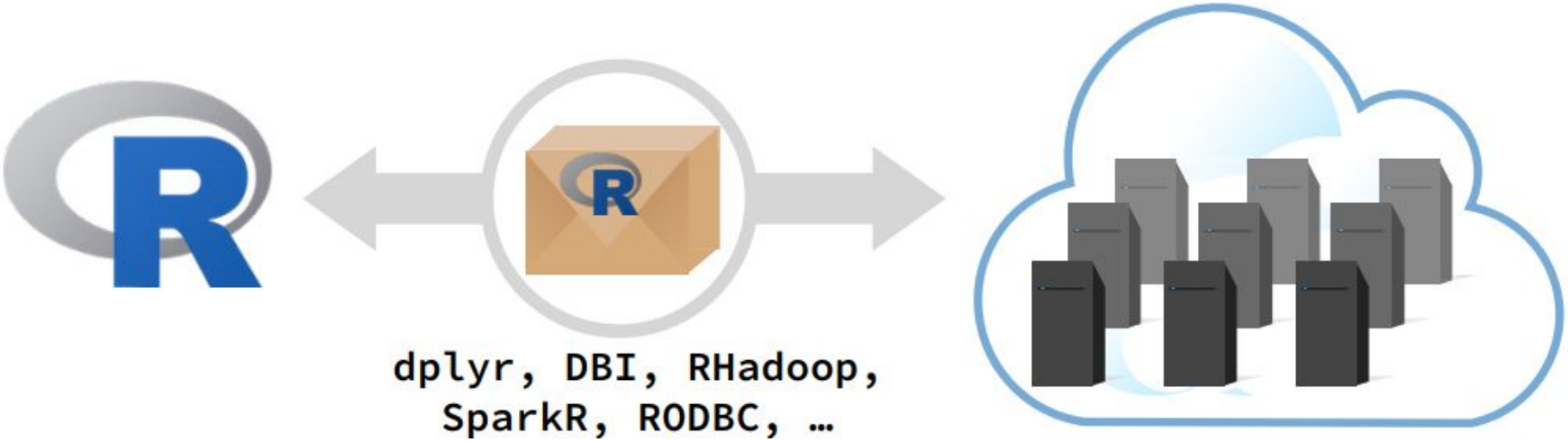
1. Pasar los subconjuntos de datos del warehouse a R.
2. Transformar el código R, pasarlo al warehouse.



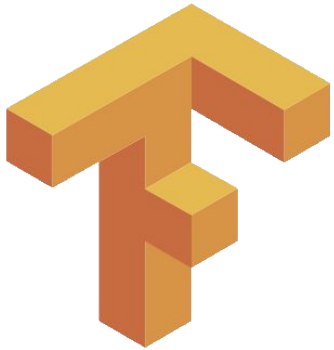
Estrategia General:

Guardar datos en un data warehouse

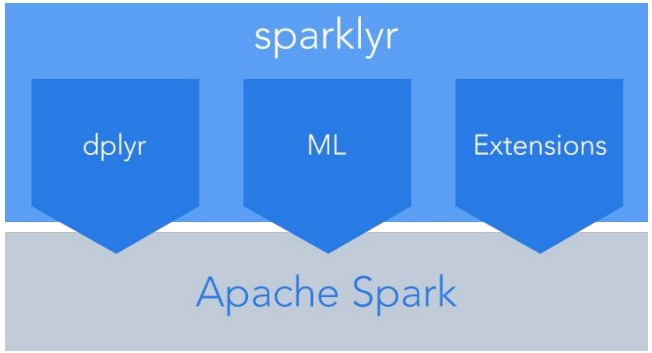
1. Pasar los subconjuntos de datos del warehouse a R.
2. Transformar el código R, pasarlo al warehouse.

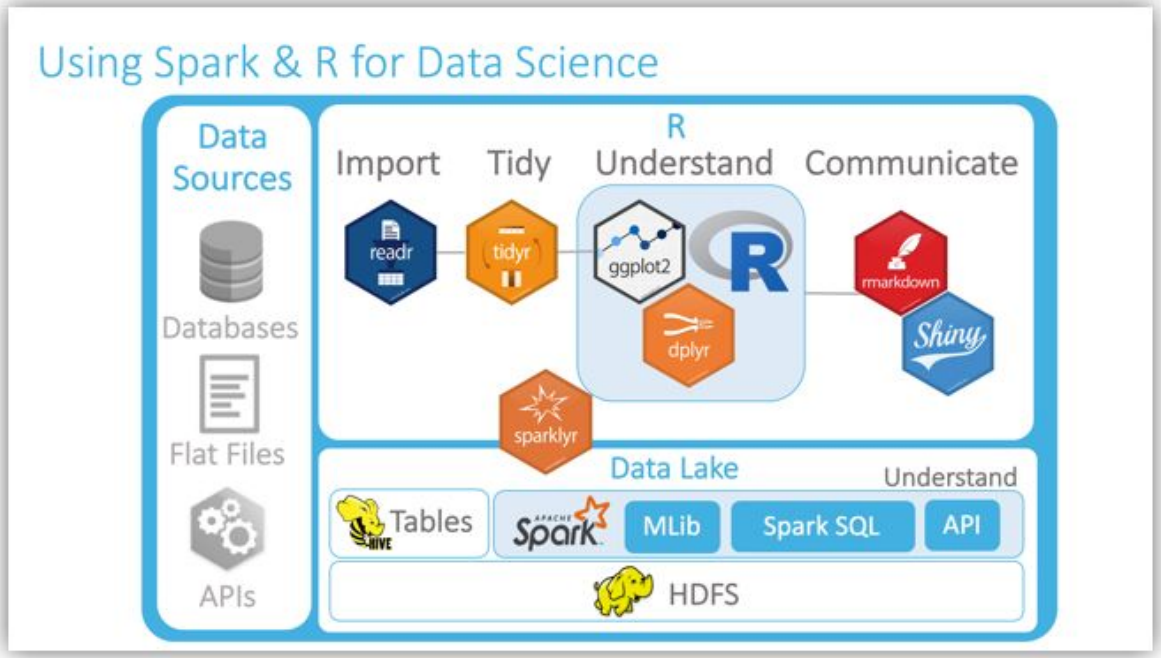



APACHE
Spark
TM




APACHE
Spark
TM

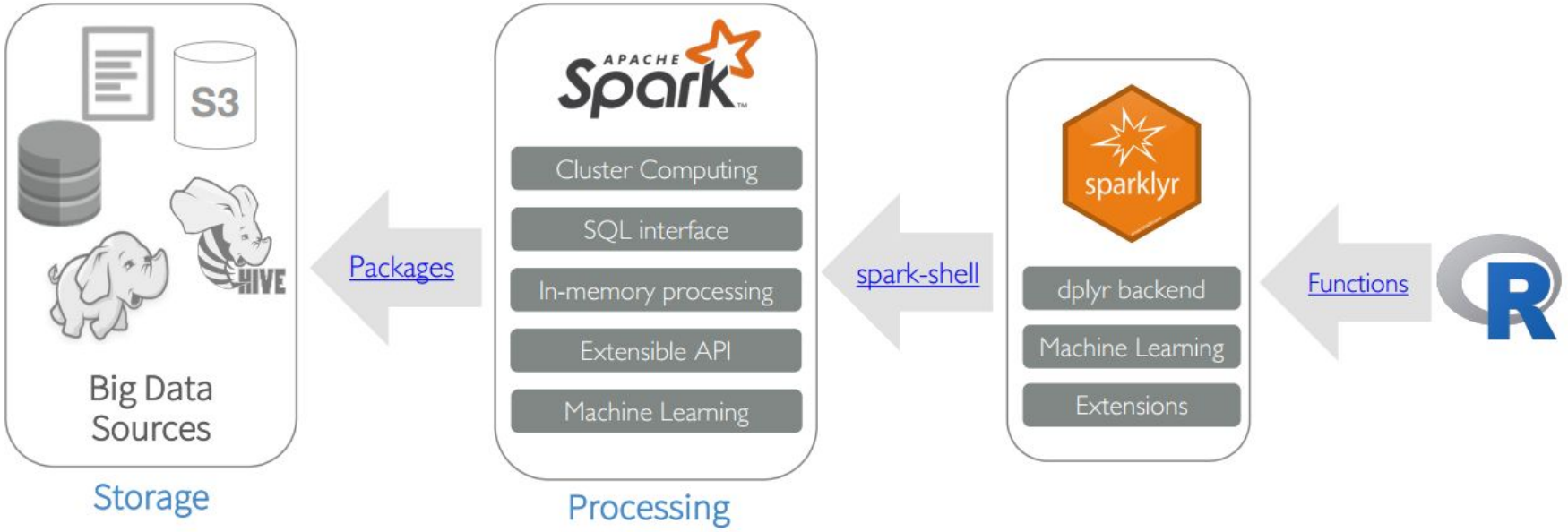

sparklyr
dplyr
ML
Extensions
Apache Spark


Using Spark & R for Data Science
Data Sources
Databases
Flat Files
APIs
R
Import
Tidy
Understand
Communicate
readr
tidyr
ggplot2
dplyr
rmarkdown
Shiny
sparklyr
Data Lake
Understand
Tables
Spark
MLib
Spark SQL
API
HDFS

Sparklyr

Utiliza a spark-shell para obtener acceso a Spark. Además de proporcionando acceso SQL, sparklyr también abre el acceso a las funciones API de Spark como Feature Transformers, modelos de Learning Machine y otros.

Sparklyr Una Interface para Spark



Leyendo Datos de Spark

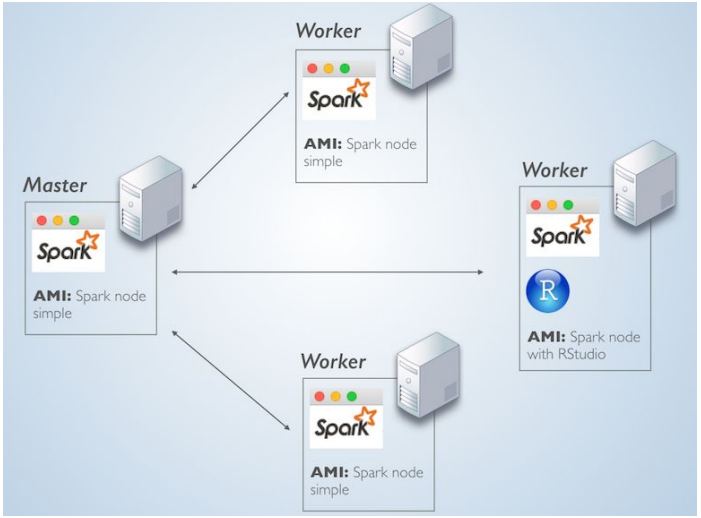
Option 1 – Read data live

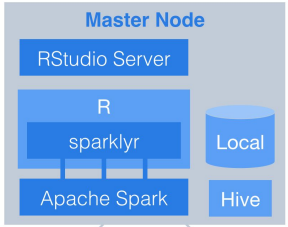
Spark will retrieve data from the source every time a new request from R is sent

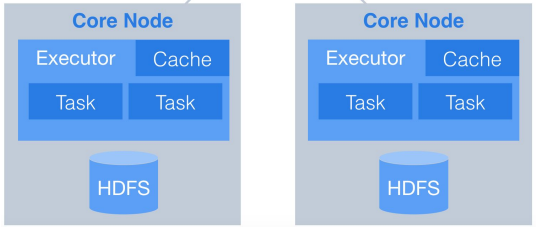
Option 2 – Cache data in Spark

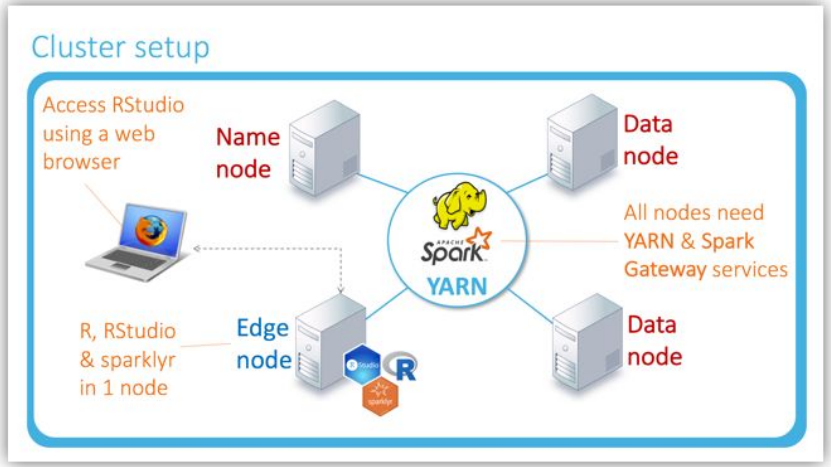
Spark will retain data in-memory while the session is active

Prueba de Concepto.









https://spark.rstudio.com

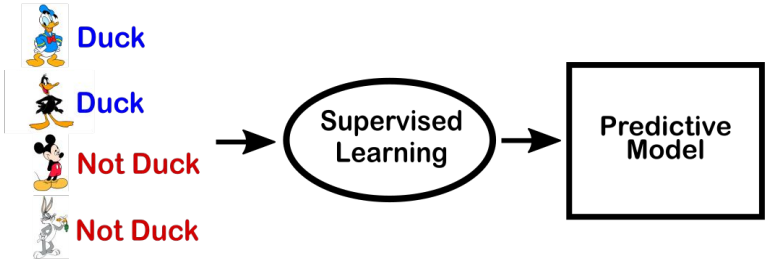
Bigrquery

Prueba de Concepto

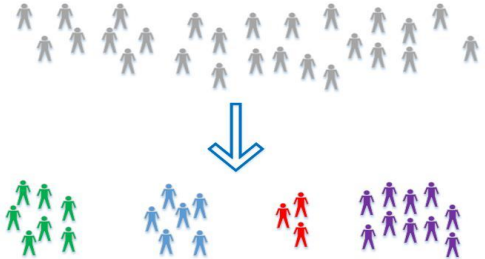
Define: Machine Learning

Tipos de Machine Learning

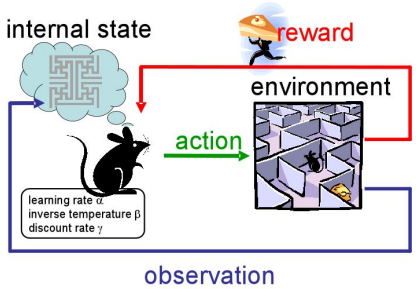
Aprendizaje supervisado



Aprendizaje no supervisado



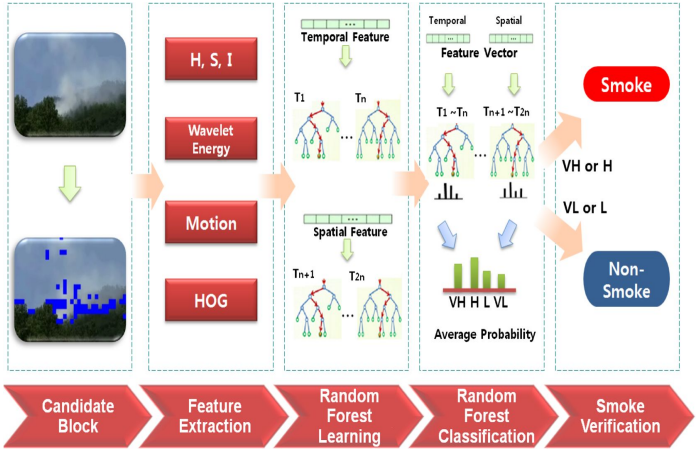
Aprendizaje por refuerzo



Algoritmos más utilizados

- Regresión Lineal
- Regresión Logística
- Árboles de Decisión
- Random Forest
- SVM o Máquinas de soporte vectorial.
- KNN o K vecinos más cercanos.
- K-means
- Redes Neuronales

Ej:

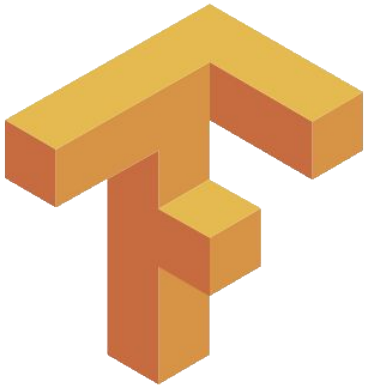


Paquetes de R para Machine Learning

https://cran.r-project.org/web/views/MachineLearning.html

Pasos para construir un Modelo de Machine Learning

- Recolectar los datos
- Preprocesar los datos
- Explorar los datos
- Seleccionar algoritmo
- Entrenar algoritmo
- Evaluar el algoritmo
- Utilizar el modelo



Tensor Flow

Deep Learning

TensorFlow con R

https://tensorflow.rstudio.com/

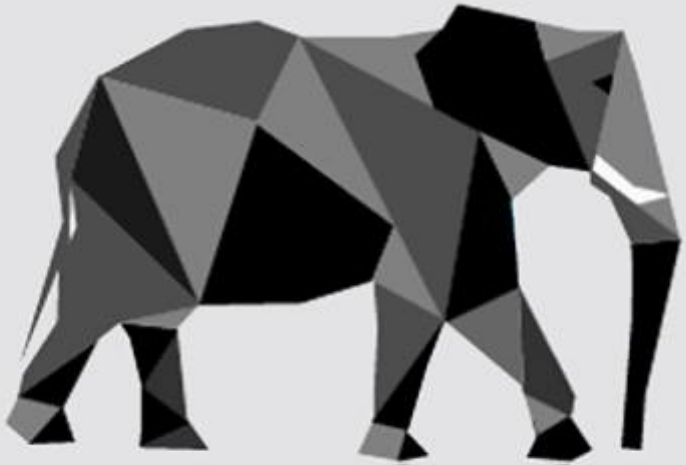
http://playground.tensorflow.org/

Prueba de Concepto

Más Info:

http://camiloherrera.co/

http://42data.co/



42Data

Turn data into
smart information