






Big Data: Una nueva oportunidad de desarrollo

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Big Data.

Descargue la última versión de este documento de:
<https://github.com/jdvelasq/Lecture-notes-on-analytics/blob/master/intersoftware-big-data.pdf>

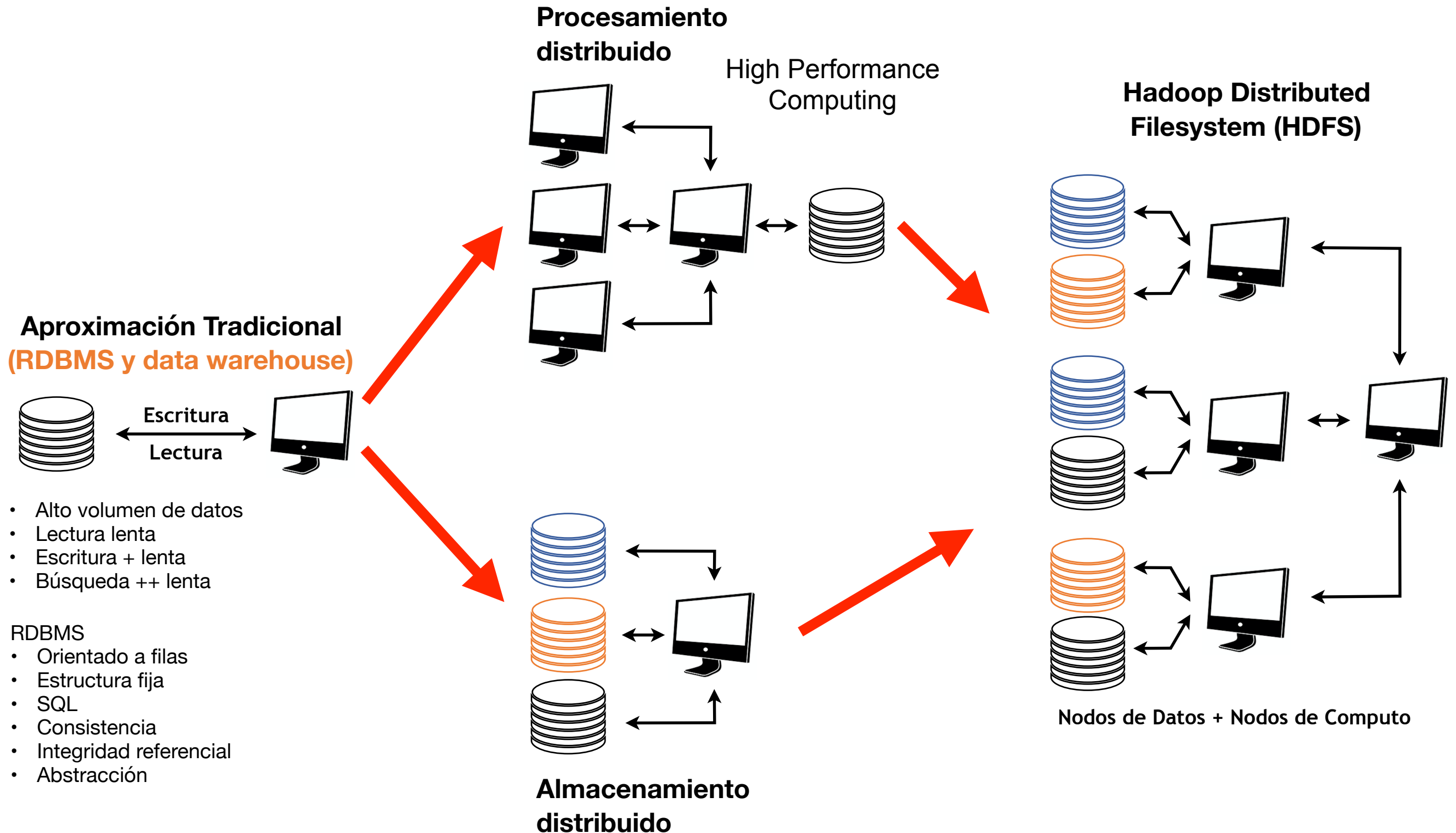
JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD
Profesor Titular
Departamento de Ciencias de la Computación y la Decisión
Facultad de Minas
Universidad Nacional de Colombia, Sede Medellín

 jdvelasq@unal.edu.co
 [@jdvelasquezh](https://twitter.com/jdvelasquezh)
 <https://github.com/jdvelasq>
 <https://goo.gl/prkjAq>
 <https://goo.gl/vXH8jy>

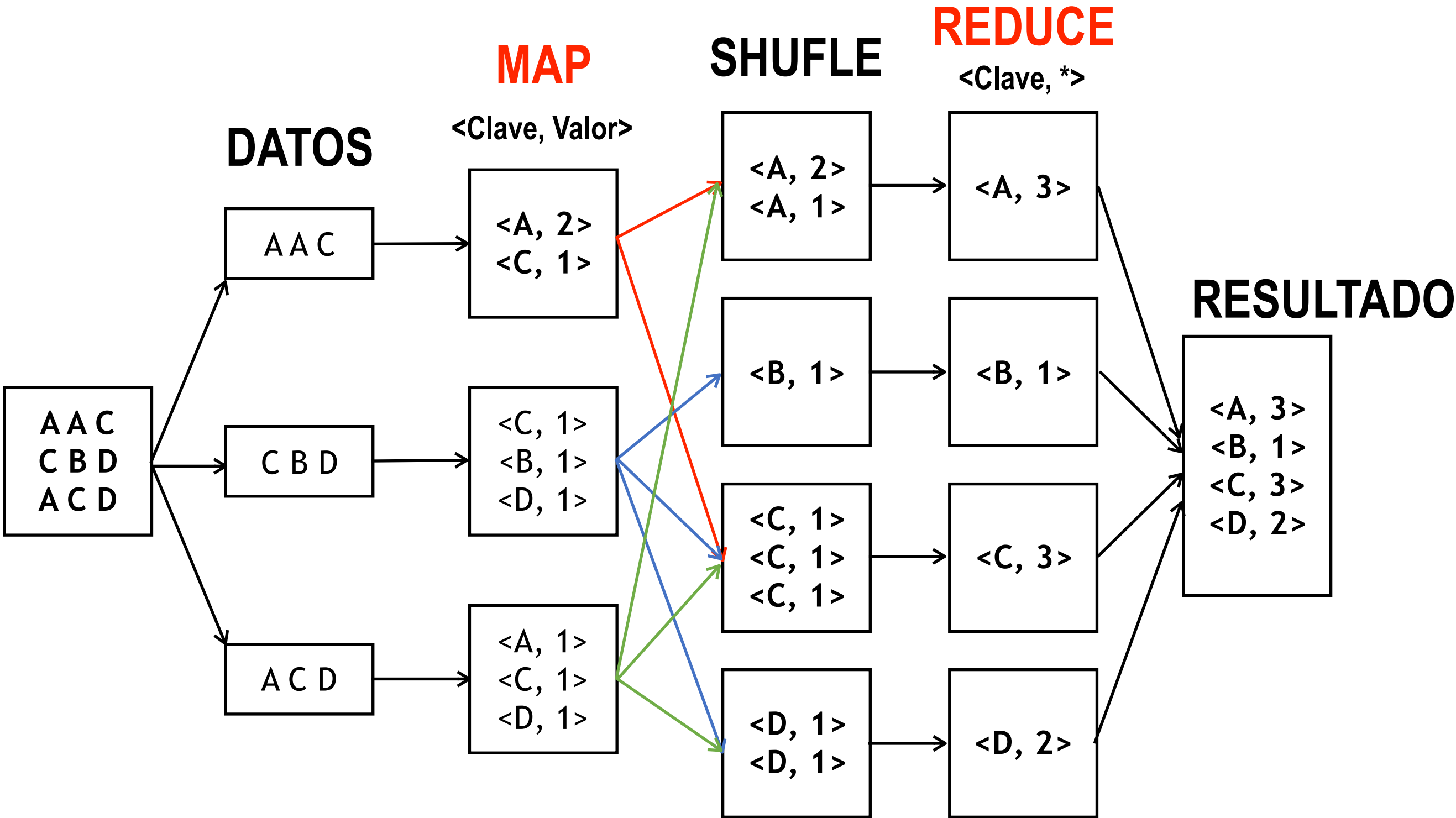
Big Data &
Data Analytics

Categoría A,
Convocatoria 781 de 2017

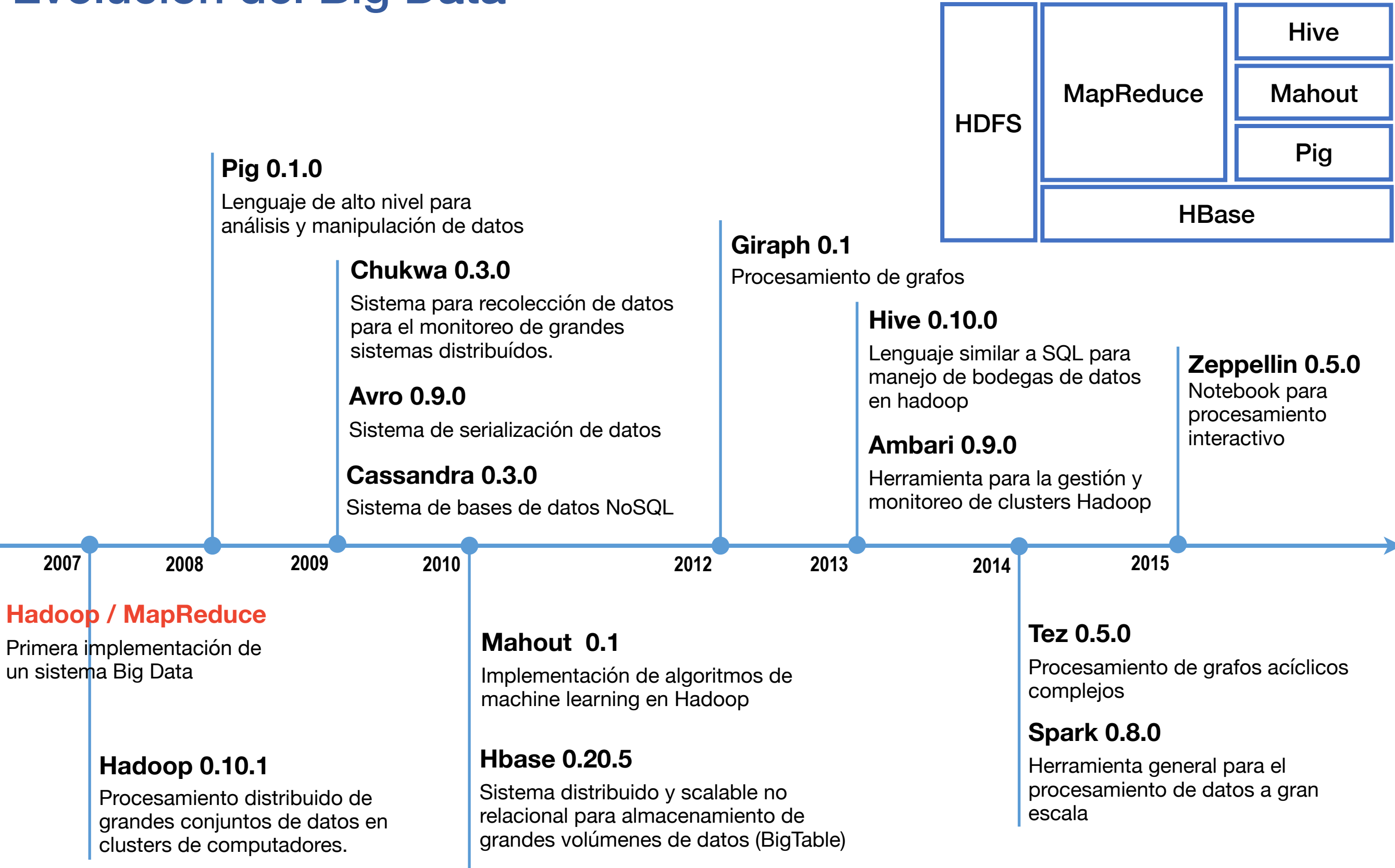
Hadoop / MapReduce (2005)



Hadoop / MapReduce (2005)



Evolución del Big Data



Pig Latin (2008)

CROSS
EXPLAIN
FILTER
FOREACH
GENERATE
GROUP
ILLUSTRATE
JOIN
LIMIT
LOAD
ORDER
STREAM
SPLIT
STORE
SET
QUIT

Lenguaje similar al SQL para el análisis de grandes volúmenes de datos en Hadoop representados como flujos de datos.

Ejemplo de Pig

```
records = LOAD 'sample.txt' AS (year:chararray, temperature:int, quality:int);  
filtered_records = FILTER records BY temperature;  
grouped_records = GROUP filtered_records BY year;  
max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);  
DUMP max_temp;
```

NoSQL (2009)

Datos tabulares

KEY	Fecha	Planta	Generación
001	2017-10-01	Jaguas	100.2
002	2017-10-01	Playas	23.1
003	2017-10-01	Guatape	130.1

Document (JSON/XML)

```
[
  {
    Fecha:2017-10-01,
    Planta:Jaguas,
    Generación: 100.2
  },{
    Fecha:2017-10-01,
    Planta:Playas,
    Generación:23.1,
  },{
    Fecha:2017-10-01,
    Planta:Guatapé,
    Generación:130.1
  }
]
```

Pares <clave, valor>

Tabla001.Fecha=2017-10-01
Tabla001.Planta=Jaguas
Tabla001.Generación=100.2
Tabla002.Fecha=2017-10-01
Tabla002.Planta=Playas
Tabla002.Generación=23.1
Tabla003.Fecha=2017-10-01
Tabla003.Planta=Guatapé
Tabla003.Generación=130.1

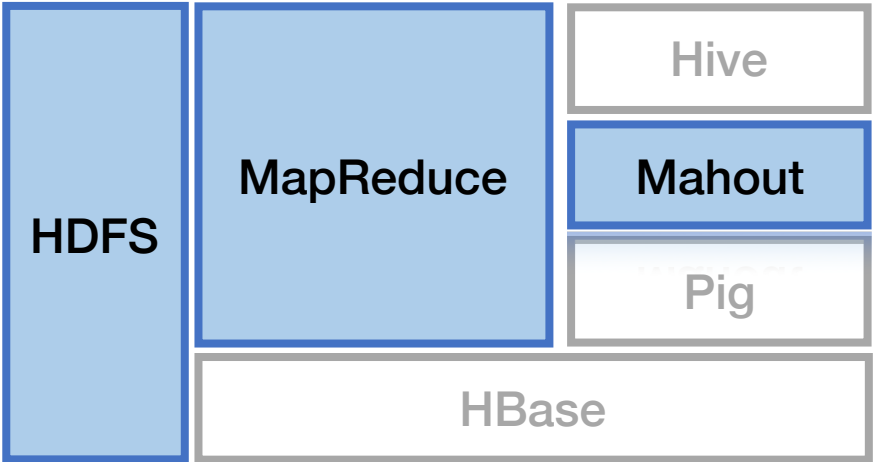
Sistema orientado a filas

001:2017-10-01,Jaguas,100.2
002:2017-10-01,Playas,23.1
003:2017-10-01,Guatape,130.1

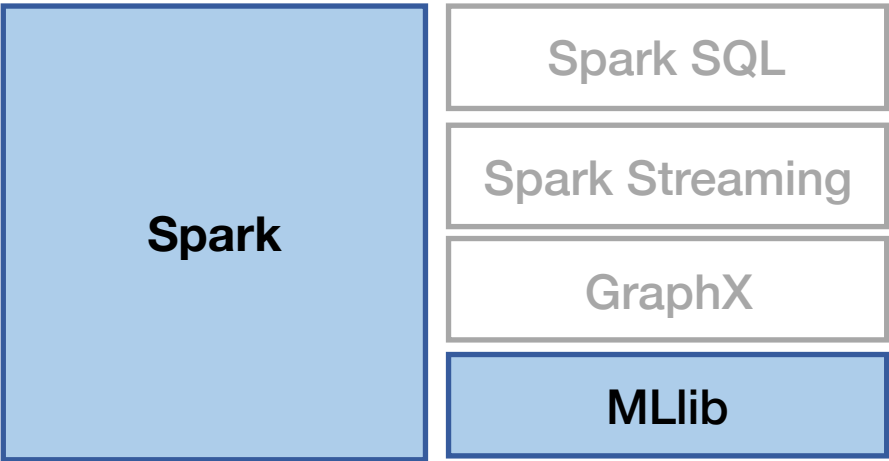
Column family database

001:{Fecha:2017-10-01, Planta:Jaguas, Generación:100.2}
002:{Fecha:2017-10-01, Planta:Playas, Generación:23.1}
003:{Fecha:2017-10-01, Planta:Guatapé, Generación:130.1}

Big Data Analytics (2011)



Apache Mahout
Implementación en Map/Reduce (Java y otros) de los algoritmos de aprendizaje estadístico y aprendizaje de máquinas



Spark's MLlib
Implementación en Spark de los algoritmos de aprendizaje estadístico y aprendizaje de máquinas

{
Java
Scala
Python
R

- Estadística básica
- Clasificación y regresión
- Filtrado colaborativo
- Agrupamiento
- Reducción de dimensiones
- Extracción de características
- Minería de patrones frecuentes
- Métricas de evaluación
- Exportación de modelos
- Optimización



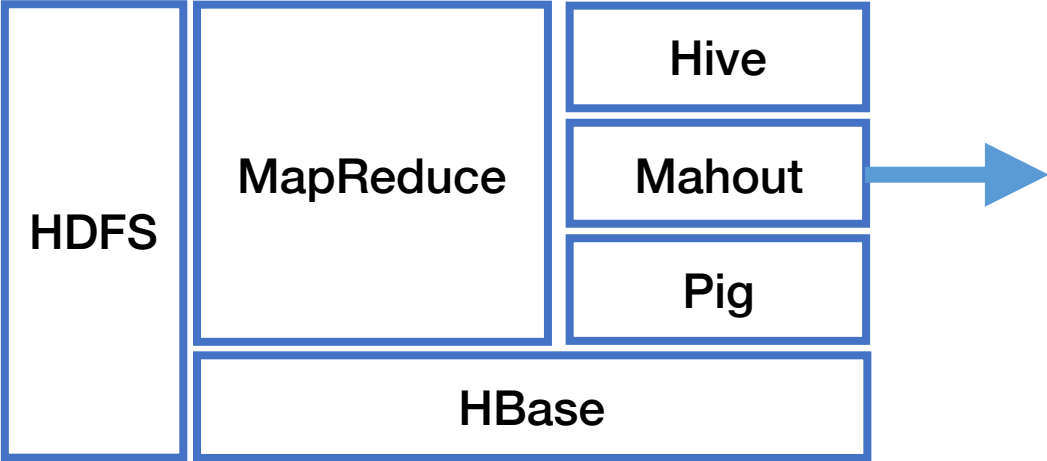
{
Computación de alto desempeño
Deep Learning

Apache Hive (2013)

Ejemplo de Hive

```
CREATE TABLE records (year STRING, temperature INT, quality INT)  
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';  
LOAD DATA LOCAL INPATH 'sample.txt' OVERWRITE INTO TABLE records;  
SELECT year, MAX(temperature) FROM records GROUP BY year;
```

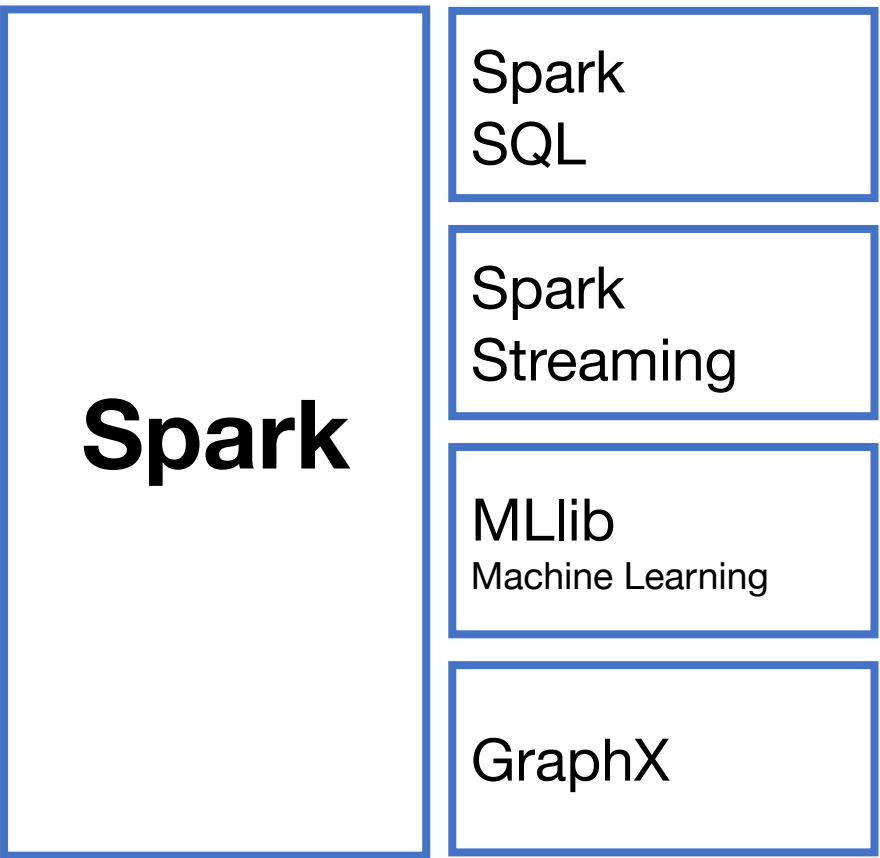
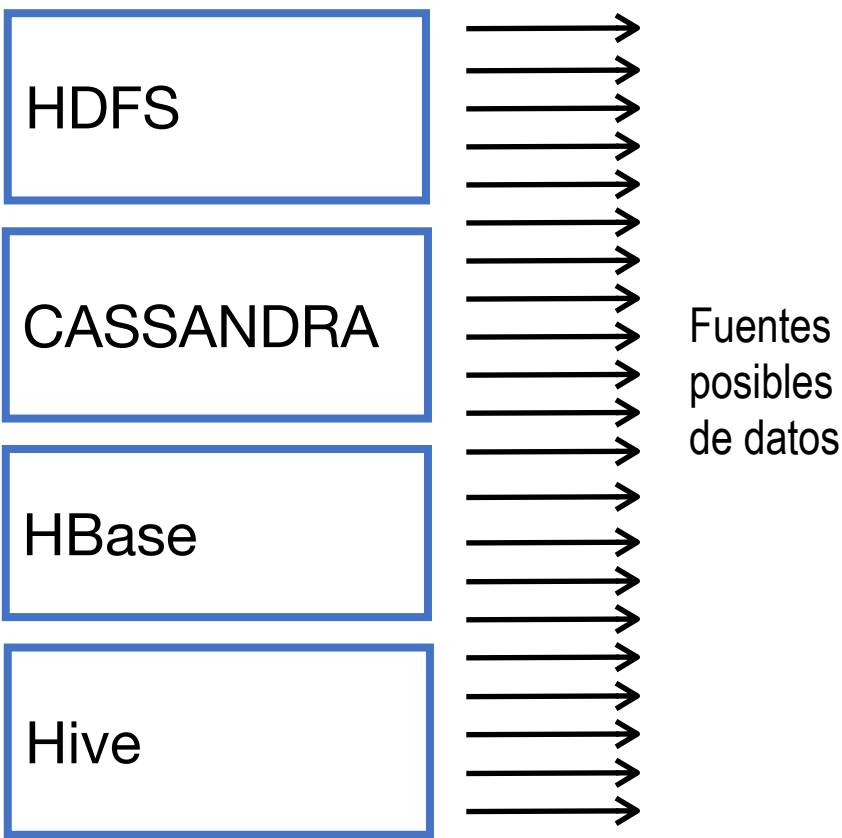

Apache Spark (2014)



Hadoop / MapReduce

Regresión logística
Regresión lineal
Clustering
Filtrado colaborativo
<http://mahout.apache.org/users/basics/algorithms.html>

RHadoop
rdfs
rmr
rhbase







Java
Scala
Python
R

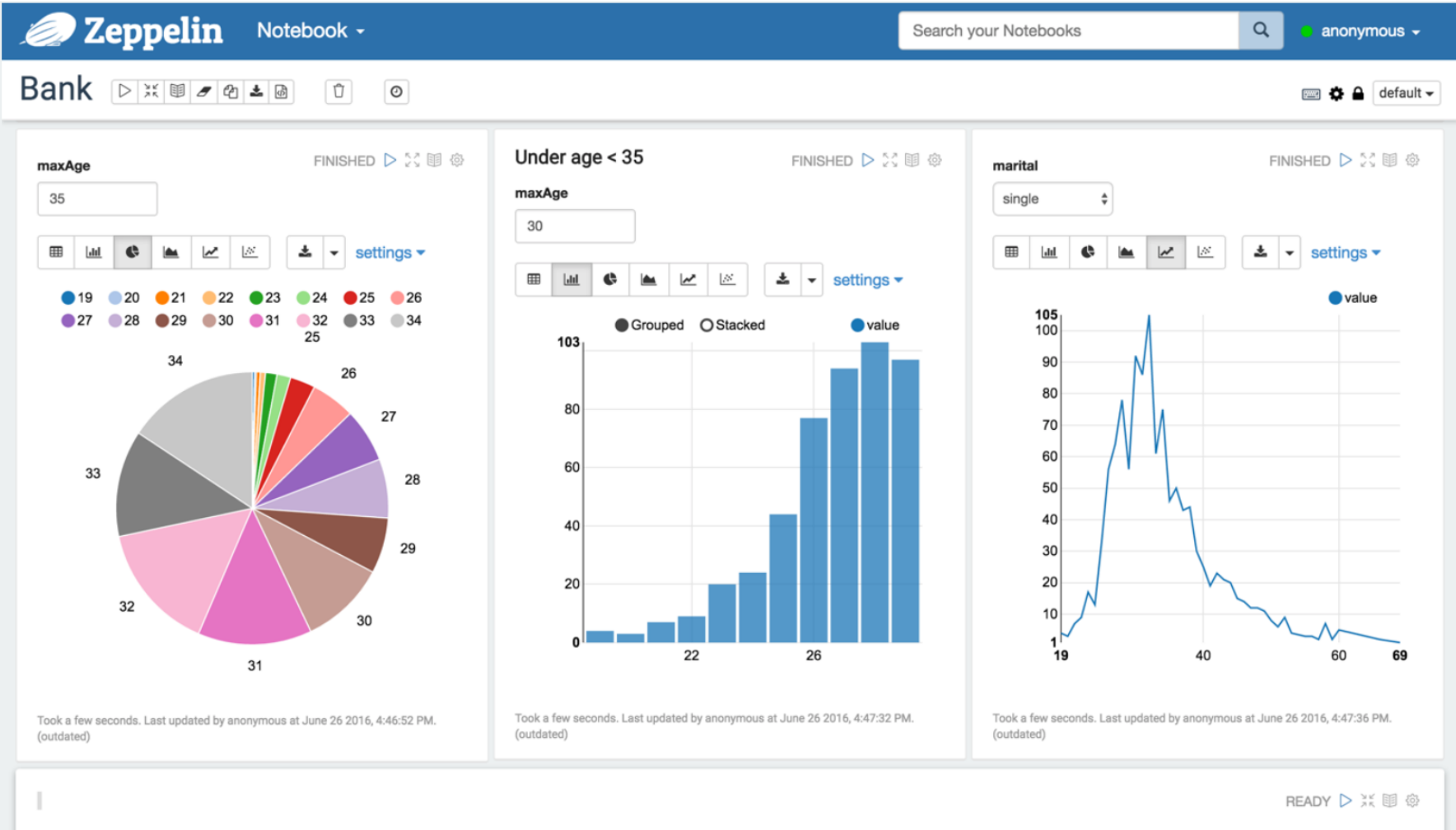
<https://spark.apache.org/mllib/>

Apache Zeppelin (2015)

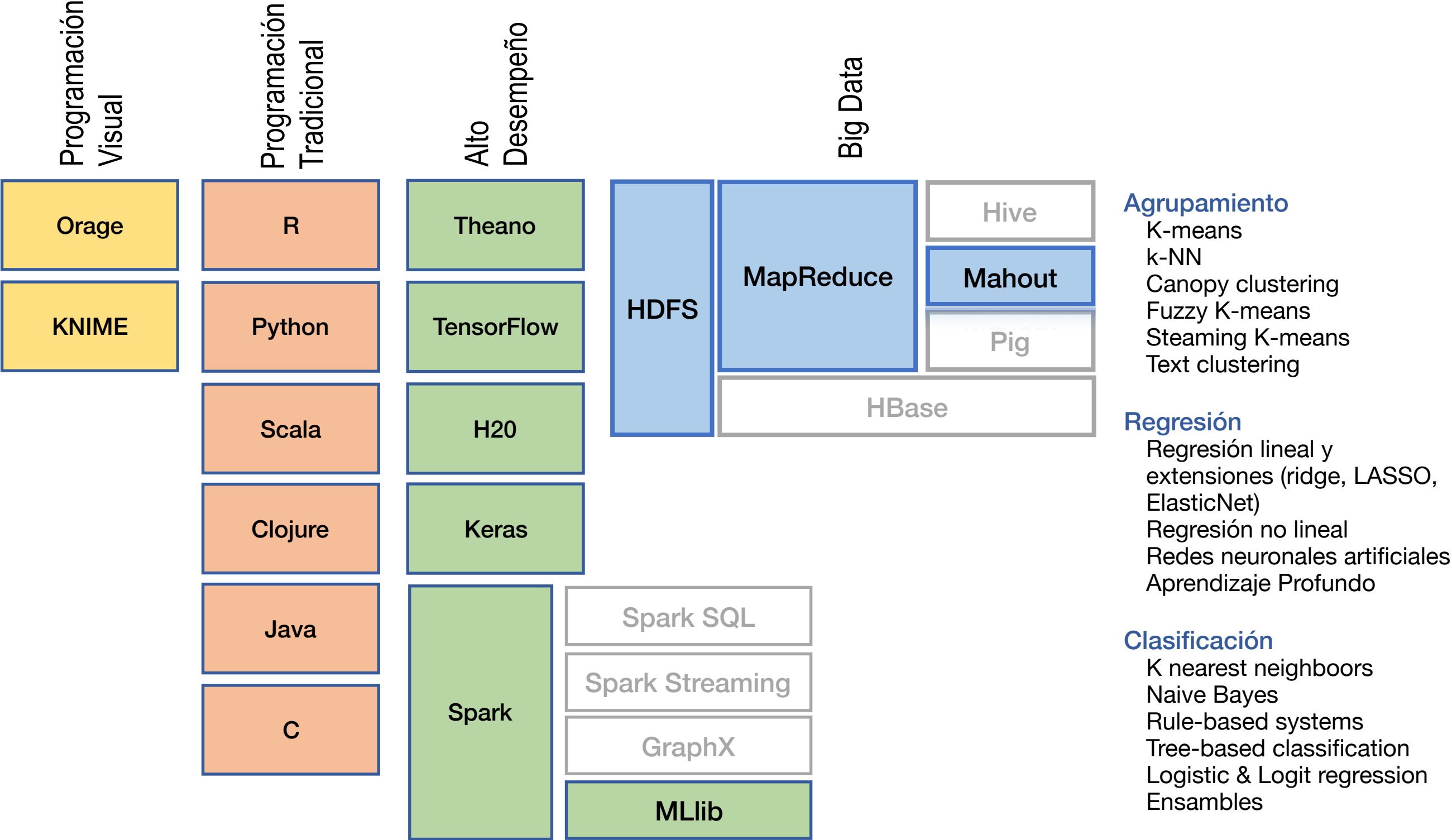
Multi-purpose Notebook

The Notebook is the place for all your needs

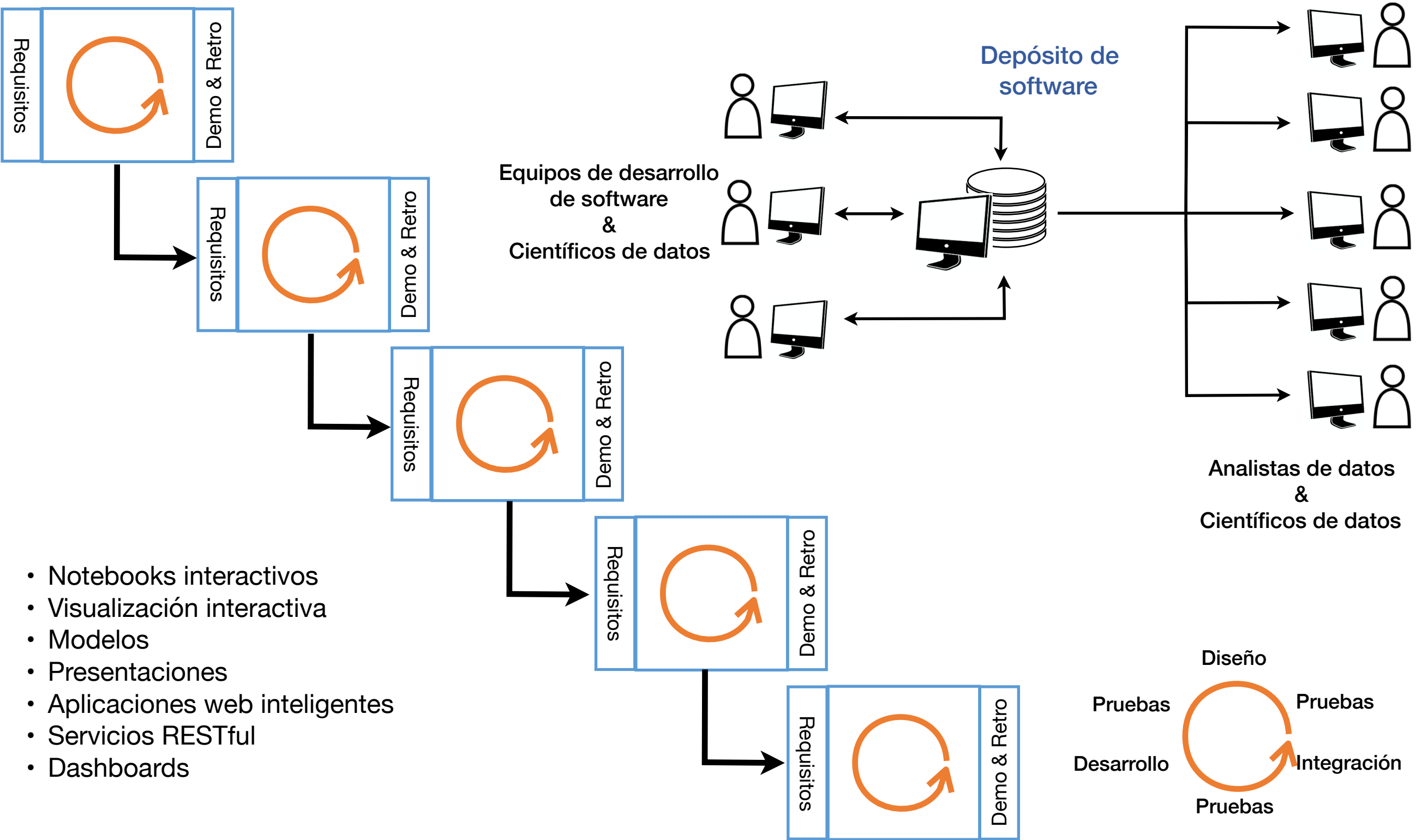
-  Data Ingestion
-  Data Discovery
-  Data Analytics
-  Data Visualization & Collaboration



Open Data Science & Modern Analytics (2018)



DataOps (2015)








- Notebooks interactivos
- Visualización interactiva
- Modelos
- Presentaciones
- Aplicaciones web inteligentes
- Servicios RESTful
- Dashboards

Big Data: Una nueva oportunidad de desarrollo

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Big Data.

Descargue la última versión de este documento de:
<https://github.com/jdvelasq/Lecture-notes-on-analytics/blob/master/intersoftware-big-data.pdf>

JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD
Profesor Titular
Departamento de Ciencias de la Computación y la Decisión
Facultad de Minas
Universidad Nacional de Colombia, Sede Medellín

 jdvelasq@unal.edu.co
 [@jdvelasquezh](https://twitter.com/jdvelasquezh)
 <https://github.com/jdvelasq>
 <https://goo.gl/prkjAq>
 <https://goo.gl/vXH8jy>

Big Data &
Data Analytics

Categoría A,
Convocatoria 781 de 2017