¿Es Big Data el uso de viejas técnicas matemáticas y estadísticas bajo una mayor capacidad de computación, o es realmente un tema nuevo en el mundo científico?

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://github.com/jdvelasq
https://goo.gl/prkjAq
https://goo.gl/vXH8jy

Big Data

Estrategia para el manejo de información caracterizada por:

- Un alto volumen de datos.
- Alta velocidad en la recepción de los datos
- Variedad: estructurada, semiestructurada, no estructurada

Analytics

Proceso científico de transformación de datos en conocimiento para mejorar el proceso de toma de decisiones [Informs].

I. Problema organizacional
II. Transformación en un problema de analytics
III. Datos
IV. Selección de la metodología
V. Desarrollo del modelo
VI. Puesta en marcha (deploy)
VII. Gestión del ciclo de vida del modelo

Analytics

Algoritmos, técnicas y metodologías

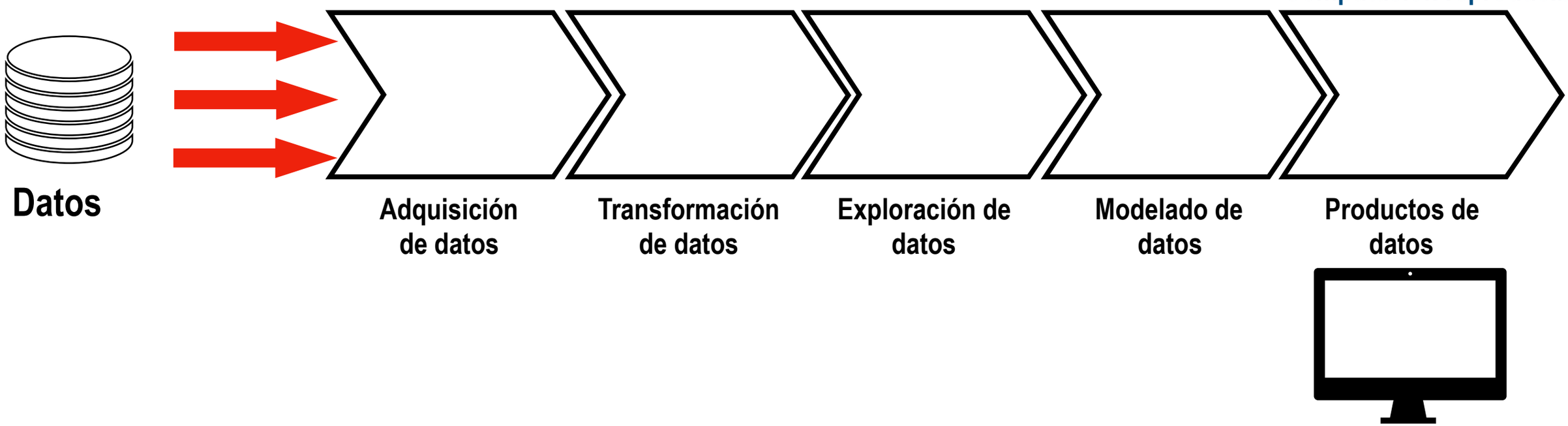
Aprendizaje estadístico
Aprendizaje de máquinas
Optimización
Simulación
Análisis de decisiones

Archivos y servicios Web
Hadoop & Spark
NoSQL
DW & SQL

Extracción,
Transformación,
Limpieza y Carga
(ETL)

Análisis
interactivo

Reportes automáticos
Dashboards
Computación reproducible



Infraestructura computacional

Hadoop / MapReduce

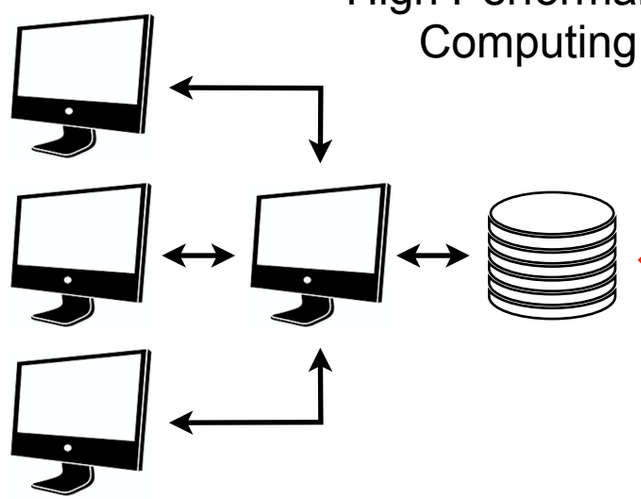
Aproximación Tradicional

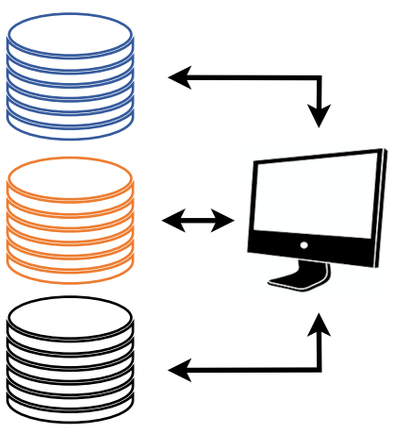
- Alto volumen de datos
- Lectura lenta
- Escritura + lenta
- Búsqueda ++ lenta

RDBMS
- Orientado a filas
- Estructura fija
- SQL
- Consistencia
- Integridad referencial
- Abstracción

Procesamiento distribuido

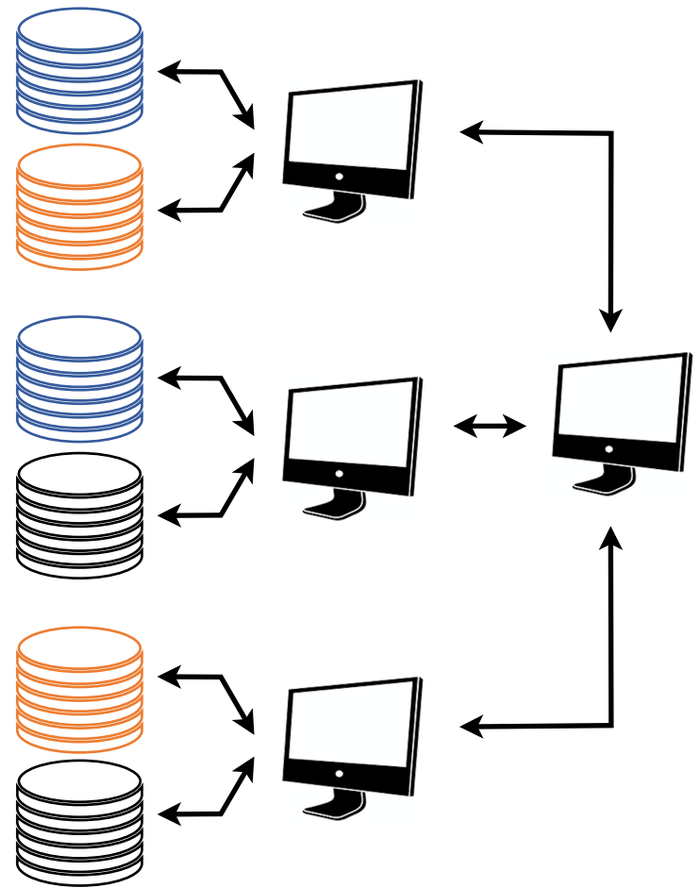
High Performance Computing





Almacenamiento distribuido

Hadoop Distributed Filesystem (HDFS)

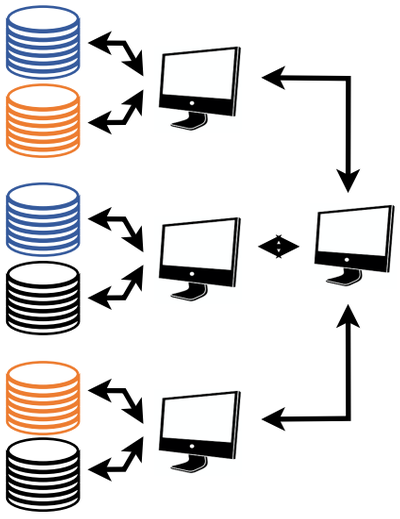


Nodos de Datos + Nodos de Computo

Infraestructura computacional

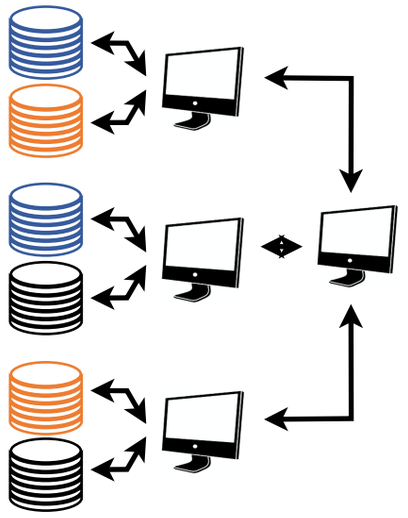
Computación local

Servidores + red + clientes



Cloud computing / utility computing

Servidores y almacenamiento en la nube + internet + clientes locales



Software as a Service (SaaS)

Software almacenado en máquinas suministradas por un tercero.

Aplicaciones accesadas vía un cliente o la Web.

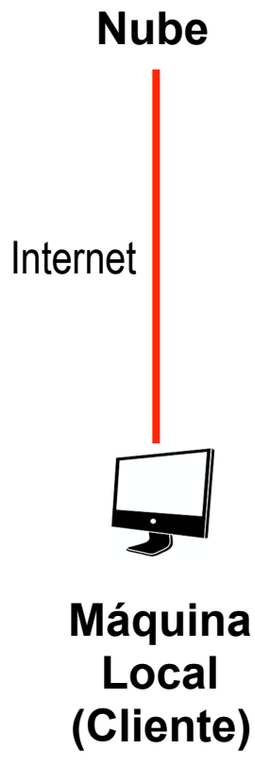
Orientado a aplicaciones de usuario final.

Platform as a Service (PaaS)

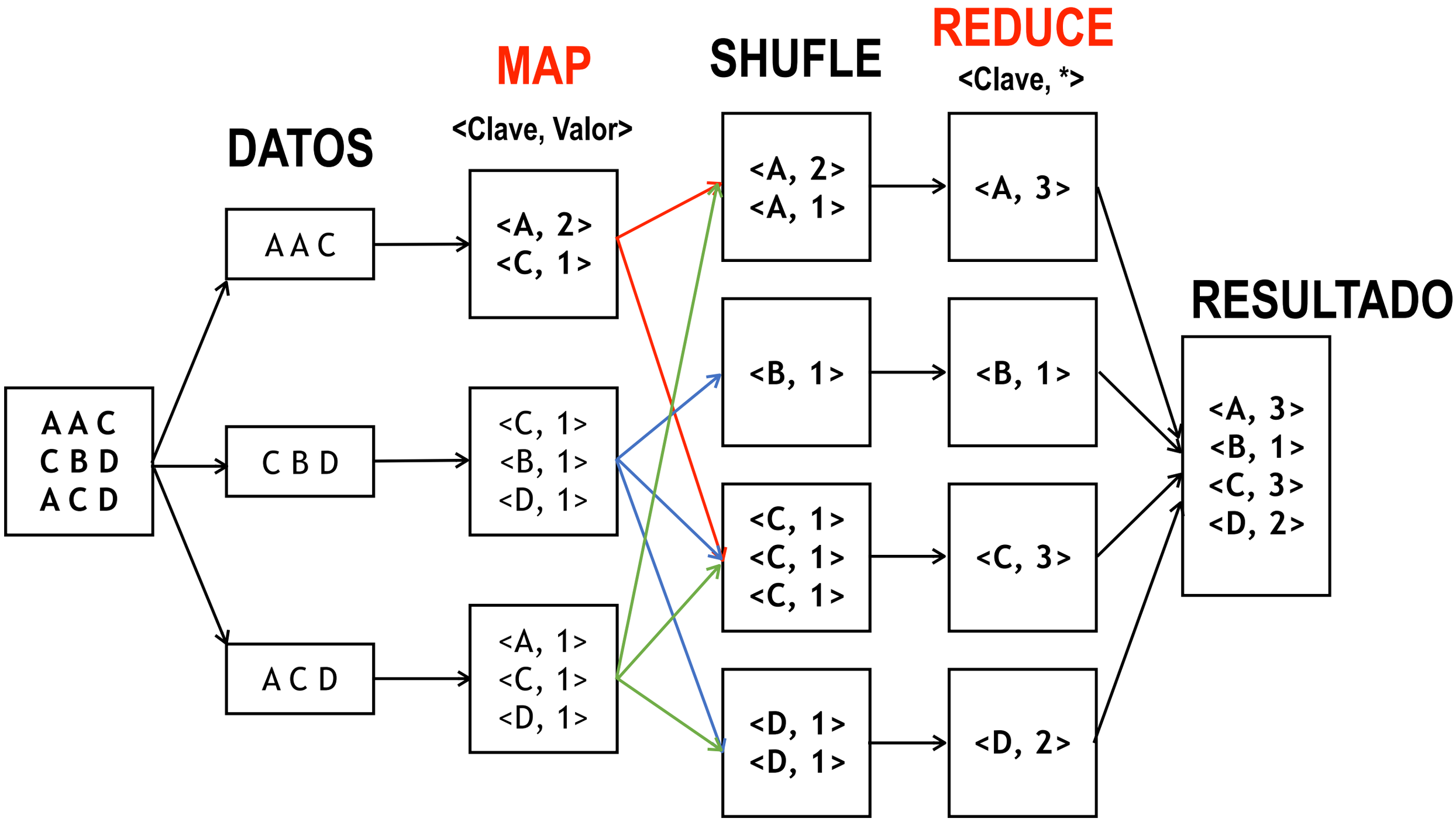
Orientado a desarrolladores.
Ambiente de desarrollo gestionado por un tercero.

Infrastructure as a Service (IaaS)

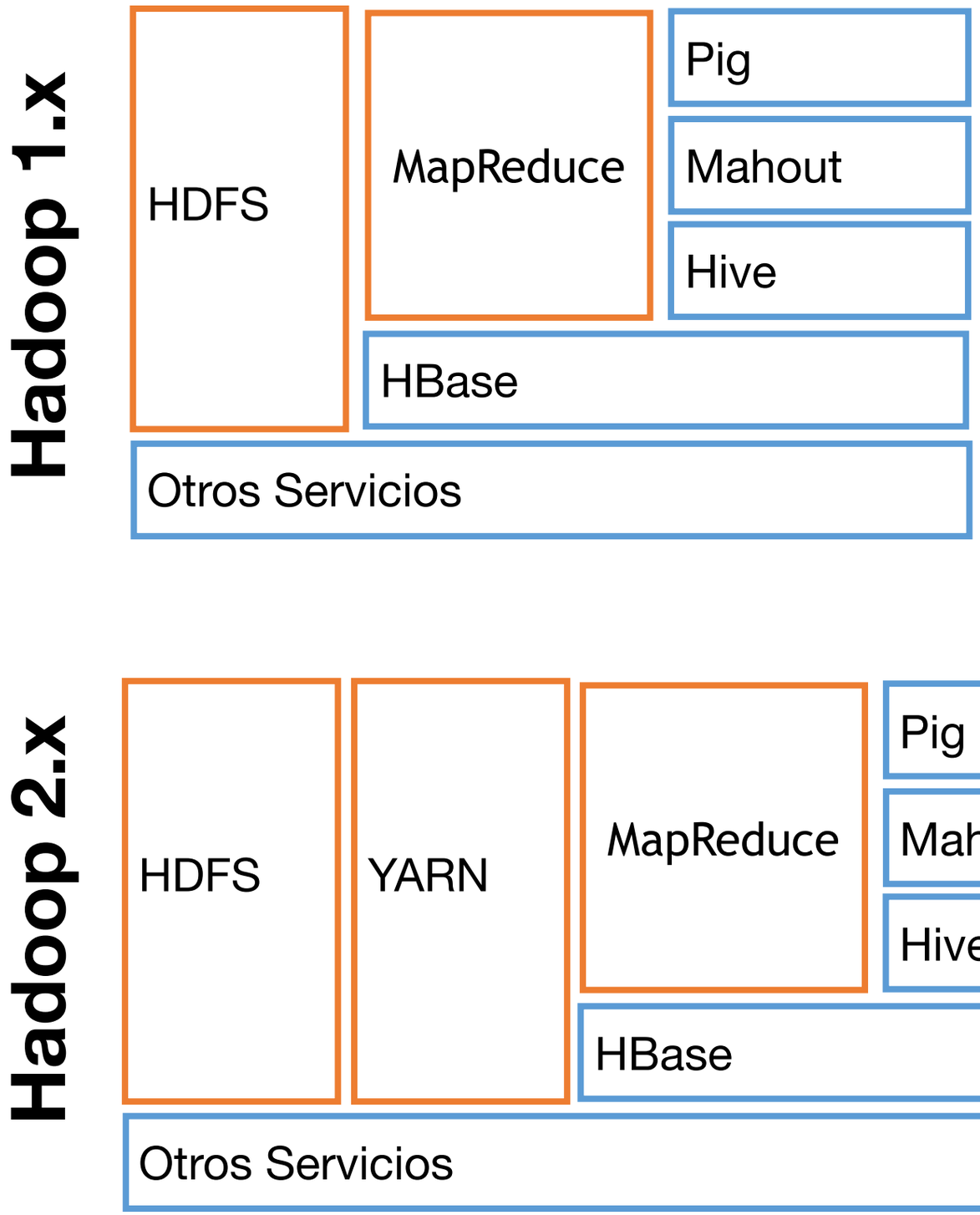
Bloques básicos para construcción de ambientes manejados por un tercero
Capacidad de procesamiento, almacenamiento, conectividad, seguridad, etc.



MapReduce


DATOS
MAP
<Clave, Valor>
SHUFLE
REDUCE
<Clave, *>
RESULTADO
A A C
C B D
A C D
A A C
C B D
A C D
<A, 2>
<C, 1>
<C, 1>
<B, 1>
<D, 1>
<A, 1>
<C, 1>
<D, 1>
<A, 2>
<A, 1>
<B, 1>
<C, 1>
<C, 1>
<C, 1>
<D, 1>
<D, 1>
<A, 3>
<B, 1>
<C, 3>
<D, 2>
<A, 3>
<B, 1>
<C, 3>
<D, 2>
<A, 3>
<B, 1>
<C, 3>
<D, 2>

Ecosistema Apache Hadoop



HDFS
Operaciones básicas del sistema de archivos

MapReduce
Algoritmo de procesamiento (Java)

HBase
Base de datos orientada a columnas.

Pig
Lenguaje para el procesamiento de datos.

Mahout
Machine Learning usando MarReduce

Hive
Lenguaje de consultas (~SQL).

YARN
Manejo y organización de recursos del cluster

Ecosistema Apache Hadoop

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;
```

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;
```

Ecosistema Apache Hadoop

HDFS
Operaciones básicas del sistema de archivos

MapReduce
Algoritmo de procesamiento (Java)

HBase
Base de datos orientada a columnas.

Pig
Lenguaje para el procesamiento de datos.

Mahout
Machine Learning usando MarReduce

Hive
Lenguaje de consultas (~SQL).

YARN
Manejo y organización de recursos del cluster

Sqoop

ZoopKeeper

Cassandra

Avro

Zeppelin

Oozie

Tez

Flume

Ambary

Kafka

Storm

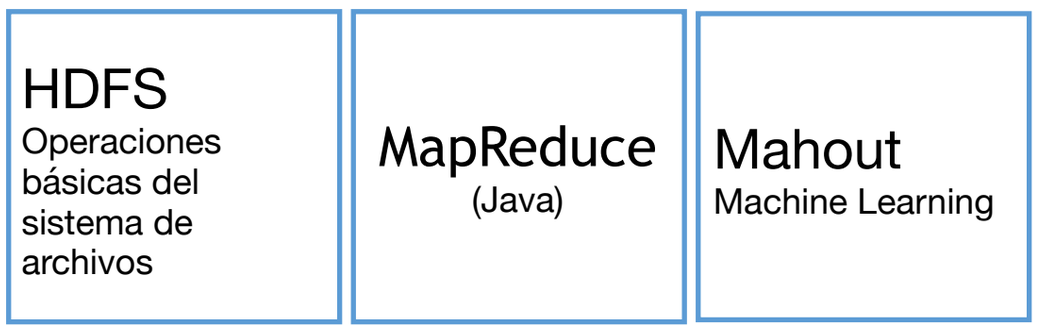
Acummulo

Solr

Knox

HAWQ

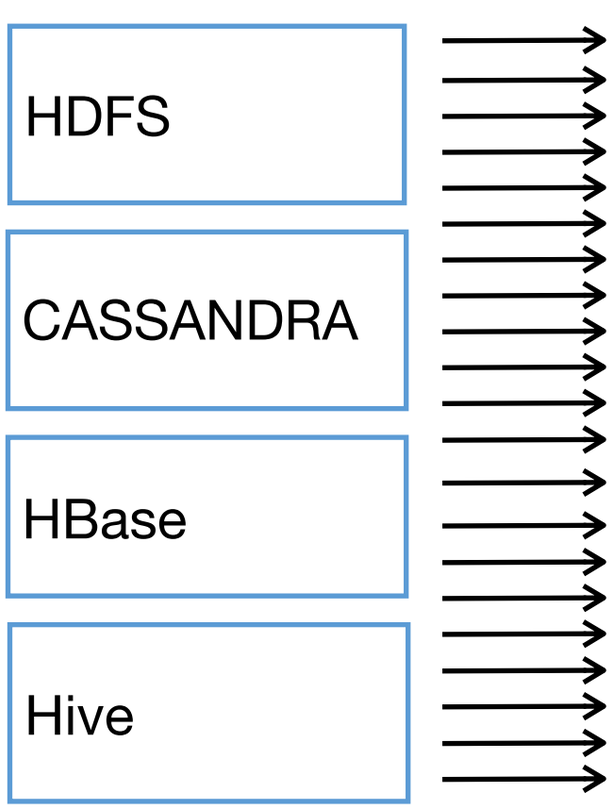
Mahout, MLib, R y Python



Regresión logística
Regresión lineal
Clustering
Filtrado colaborativo

RHadoop
rdfs
rmr
rhbase

http://mahout.apache.org/users/basics/algorithms.html



Java
Scala
Python
R

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

Mahout vs MLib

Collaborative Filtering *with CLI drivers*

User-Based Collaborative Filtering

Item-Based Collaborative Filtering

Matrix Factorization with ALS

Matrix Factorization with ALS on Implicit Feedback

Weighted Matrix Factorization, SVD++

Classification *with CLI drivers*

Logistic Regression - trained via SGD

Naive Bayes / Complementary Naive Bayes

Hidden Markov Models

Clustering *with CLI drivers*

Canopy Clustering

k-Means Clustering

Fuzzy k-Means

Streaming k-Means

Spectral Clustering

Dimensionality Reduction *note: most scala*
reduction algorithms are available through th
Core Library for all engines
Singular Value Decomposition

Lanczos Algorithm

Stochastic SVD

PCA (via Stochastic SVD)

QR Decomposition

Topic Models

Latent Dirichlet Allocation

Miscellaneous

RowSimilarityJob

Collocations

Sparse TF-IDF Vectors from Text

XML Parsing

Email Archive Parsing

Evolutionary Processes

MLib
Machine Learning

Algorithms

MLlib contains many algorithms and utilities.

ML algorithms include:

- Classification: logistic regression, naive Bayes,...
- Regression: generalized linear regression, survival regression,...
- Decision trees, random forests, and gradient-boosted trees
- Recommendation: alternating least squares (ALS)
- Clustering: K-means, Gaussian mixtures (GMMs),...
- Topic modeling: latent Dirichlet allocation (LDA)
- Frequent itemsets, association rules, and sequential pattern mining

ML workflow utilities include:

- Feature transformations: standardization, normalization, hashing,...
- ML Pipeline construction
- Model evaluation and hyper-parameter tuning
- ML persistence: saving and loading models and Pipelines

Other utilities include:

- Distributed linear algebra: SVD, PCA,...
- Statistics: summary statistics, hypothesis testing,...

Ahora la respuesta a la pregunta

Si y No, pero tal vez

Elementos teóricos y conceptuales.

Mucha teoría, técnicas y herramientas desarrolladas para un puñado de datos.

Elementos prácticos

Los resultados numéricos y tiempos de cómputo son afectados por el gran volumen de datos.

Los nuevos algoritmos deben aprovechar la infraestructura computacional existente.

Elementos de la práctica profesional

Los profesionales deben dominar las herramientas computacionales del estado del arte.

¿Es Big Data el uso de viejas técnicas matemáticas y estadísticas bajo una mayor capacidad de computación, o es realmente un tema nuevo en el mundo científico?

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://github.com/jdvelasq
https://goo.gl/prkjAq
https://goo.gl/vXH8jy