Procesamiento de datos masivos

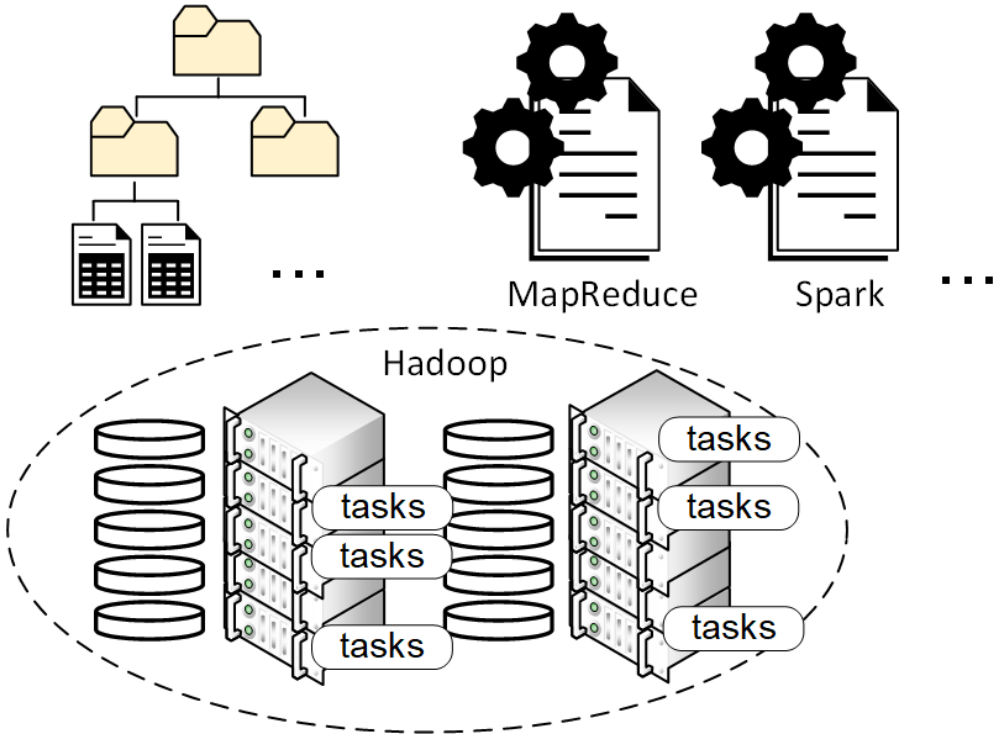
Jesús Morán

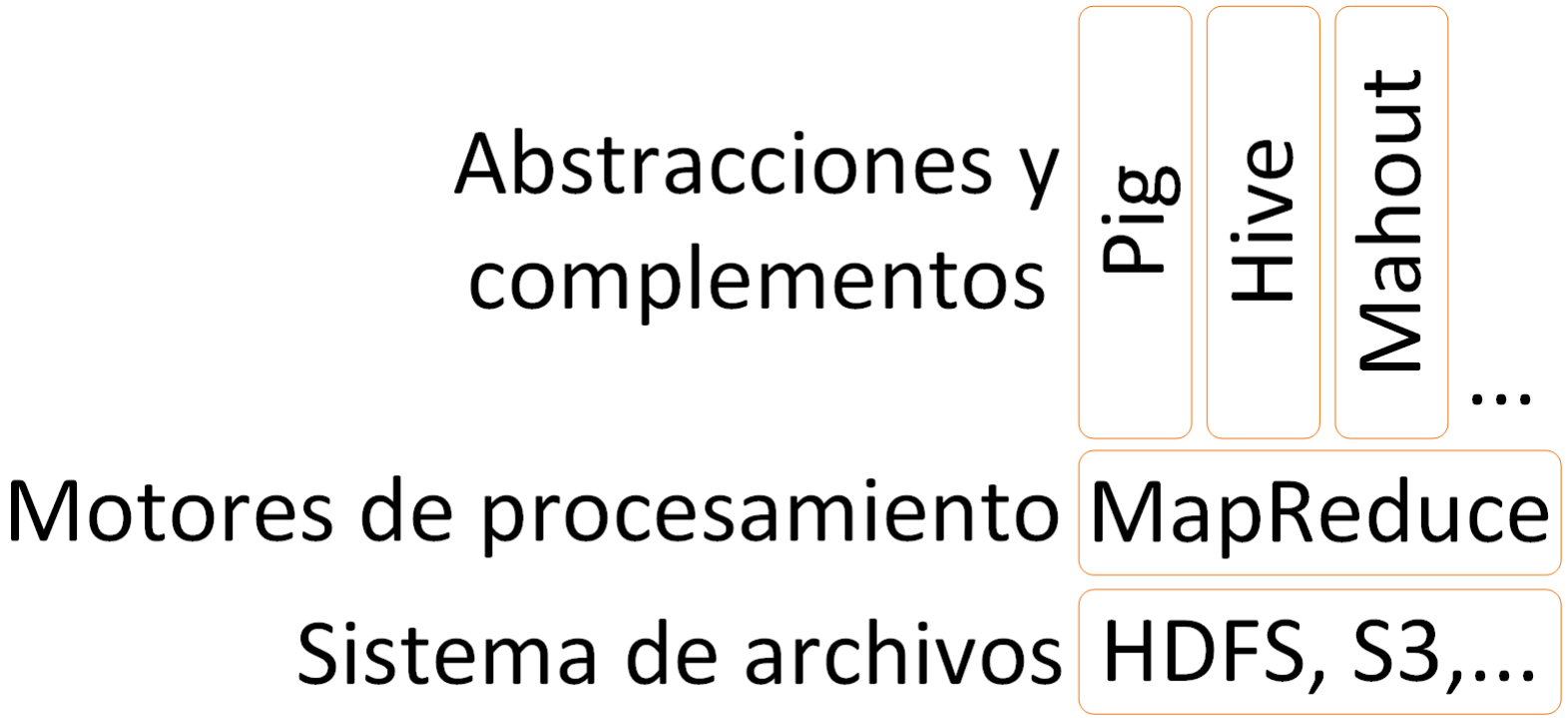
viu | Universidad Internacional de Valencia

17.01.22

- Hadoop
- Ecosistema Hadoop v1
- Ecosistema Hadoop v2
- Ecosistema Spark
- Ecosistema Flink

Hadoop



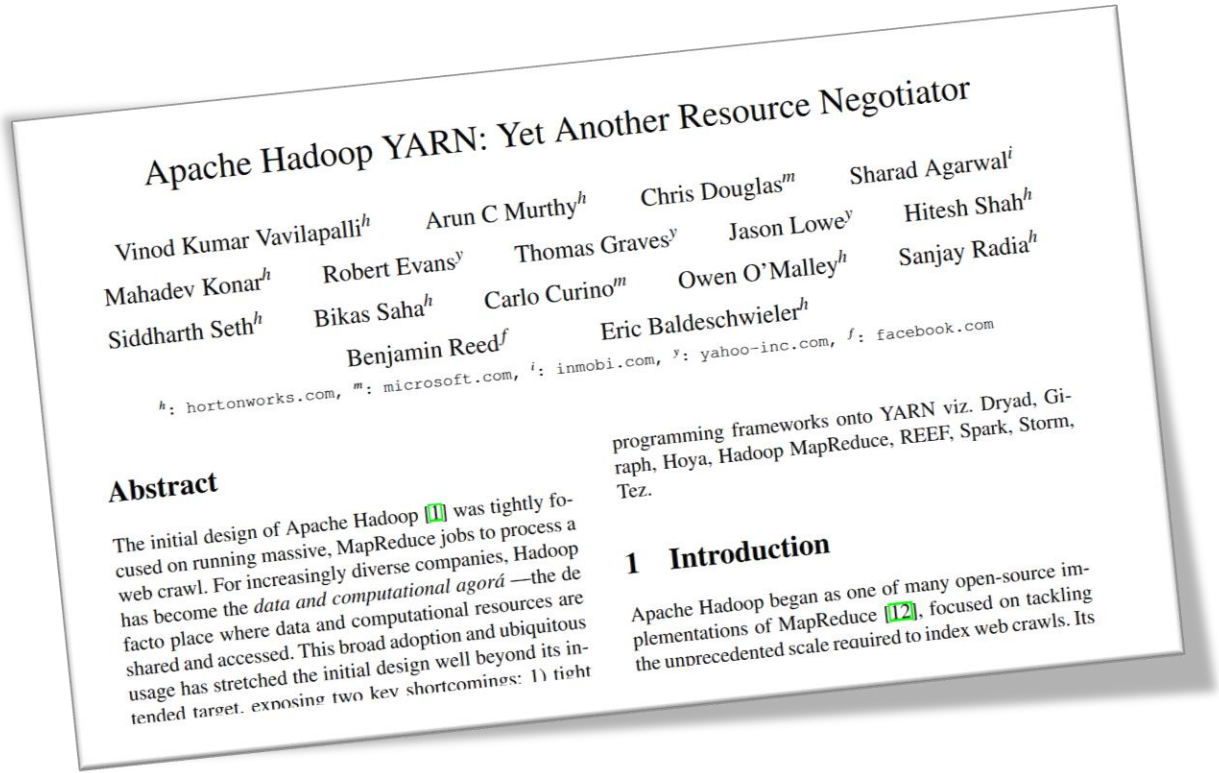


- Plataforma de análisis de datos
  - Lenguaje de alto nivel: PigLatin
    - LOAD, GROUP BY, FOREACH, STORE, …
    - http://pig.apache.org/docs/r0.17.0/basic.html
  - Se compila en jobs MapReduce
  - Ejecución: interactivo (grunt shell) y batch

- Data warehouse
  - Lenguaje de alto nivel: HiveQL (SQL-like)
    - Select, from, join, group by,…
    - Permite utilizar funciones MapReduce
    - https://cwiki.apache.org/confluence/display/Hive/GettingStarted
  - Se compila en jobs MapReduce
  - Acceso: CLI, Web y JDBC

- Librería de algoritmos Machine Learning para Big Data
  - Ejecuta jobs MapReduce, Spark,…
  - Algebra lineal
  - Clasificación
  - Clustering
  - Sistemas de recomendación
  - …

YARN



Apache Hadoop YARN: Yet Another Resource Negotiator

Vinod Kumar Vavilapalli[h] Arun C Murthy[h] Chris Douglas[m] Sharad Agarwal[i]
Mahadev Konar[h] Robert Evans[y] Thomas Graves[y] Jason Lowe[y] Hitesh Shah[h]
Siddharth Seth[h] Bikas Saha[h] Carlo Curino[m] Owen O'Malley[h] Sanjay Radia[h]
Benjamin Reed[f] Eric Baldeschwieler[h]

[h]: hortonworks.com, [m]: microsoft.com, [i]: inmobi.com, [y]: yahoo-inc.com, [f]: facebook.com

Abstract

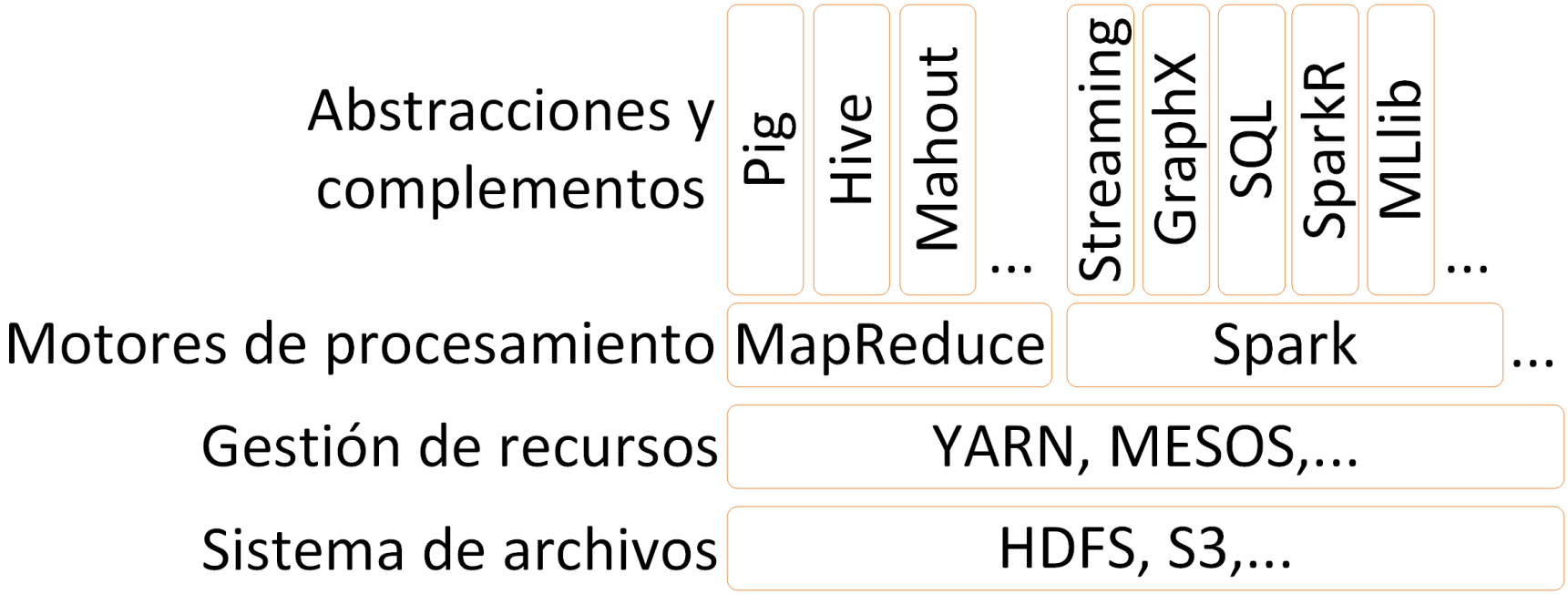
The initial design of Apache Hadoop [1] was tightly focused on running massive, MapReduce jobs to process a web crawl. For increasingly diverse companies, Hadoop has become the *data and computational agorá* —the de facto place where data and computational resources are shared and accessed. This broad adoption and ubiquitous usage has stretched the initial design well beyond its intended target, exposing two key shortcomings: 1) tight

programming frameworks onto YARN viz. Dryad, Giraph, Hoya, Hadoop MapReduce, REEF, Spark, Storm, Tez.

1 Introduction

Apache Hadoop began as one of many open-source implementations of MapReduce [12], focused on tackling the unprecedented scale required to index web crawls. Its

Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., ... & Saha, B. (2013, October). Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing* (p. 5). ACM.



- Procesamiento de datos en streaming
  - DStream: operaciones RDD en streaming
  - Structured Streaming: datos (semi)estructurados con DataFrame/Dataset al que se añaden filas dinámicamente
  - Modo de ejecución:
    - Micro-batching: batches sobre streaming
    - Modo procesamiento continuo (experimental): más streaming

- Procesamiento de grafos
  - Redes sociales, datos de internet,…
  - Grafos: vértices (VertexRDD) y nodos (EdgeRDD)
  - Operaciones sobre grafos:
    - numEdges, joinVertices, pageRank,…
    - https://spark.apache.org/docs/latest/graphx-programming-guide.html

- Análisis masivo de datos estructurados
  - Lenguaje de alto nivel: SQL o HiveQL
  - La consulta obtiene un DataFrame
    - Se pueden realizar operaciones al RDD: map, reduceByKey,….
  - Conector jdbc/odbc para ejecutar SPARK SQL

- Paquete R para análisis de datos con Spark
  - Se puede utilizar con Rstudio y otros IDEs
  - Utiliza la API DataFrame de Spark
    - Parecido a los DataFrames de R
      - SparkDataFrames se ejecuta en Spark
      - Sólo soporta algunas operaciones de los dataframes:
        - Obtener columnas, filtrar filas, agregaciones,…

- Supervisado y no supervisado
- Creación de modelos
  - Transformers para obtener las features
  - Estimators para crear modelos
  - Evaluators para evaluar los modelos
  - Pipeline

Pandas on Spark

- API pandas
- Ejecución en cluster spark

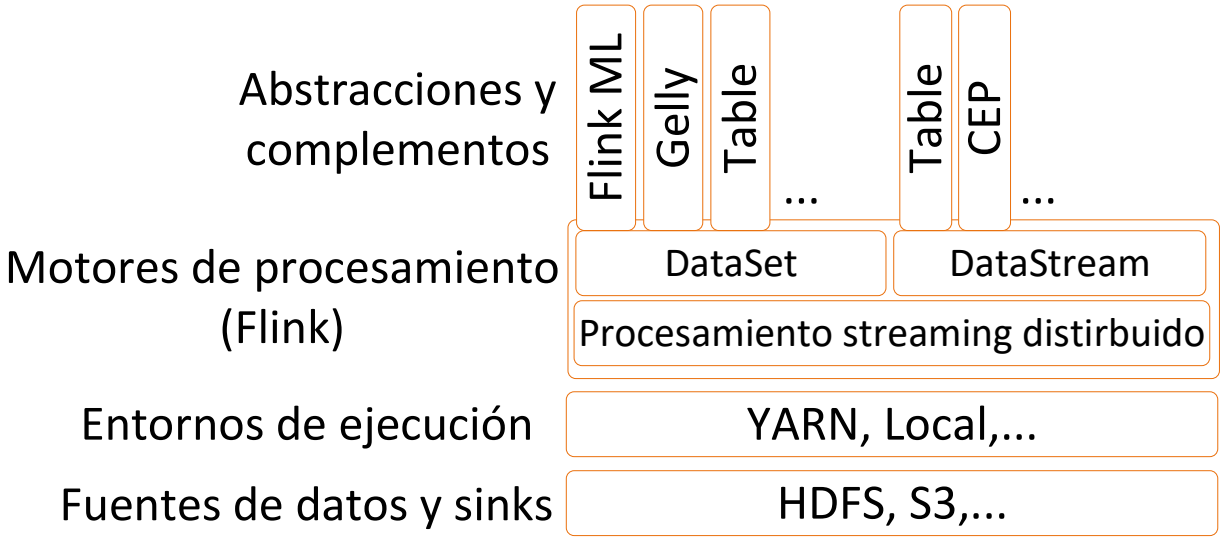
```
import pyspark.pandas as ps
```

- Disponible a partir de Spark 3.2
  - Original de Databricks (Koalas)

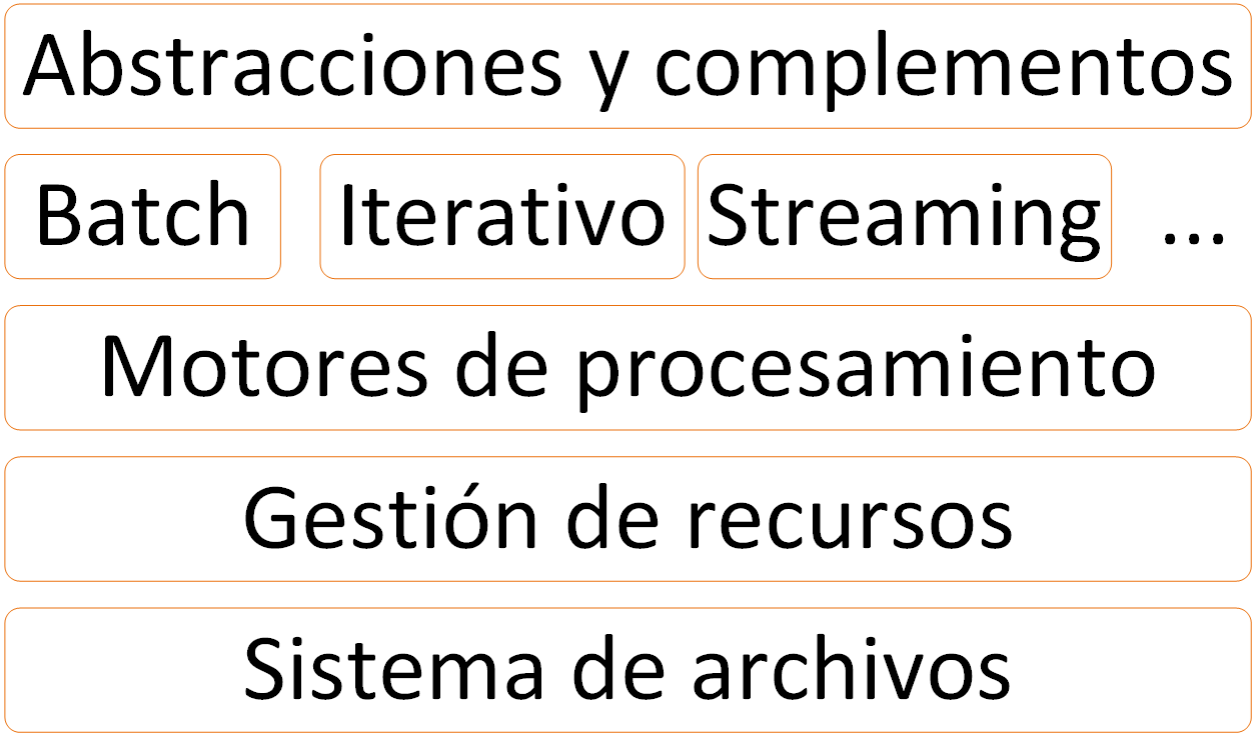
```
import databricks.koalas as ks
```

- Fácil conversión con DataFrame:
  - DataFrame -> Pandas-on-Spark: `miDF.to_pandas_on_spark()`
  - Pandas-on-Spark -> DataFrame: `miDF_pd_spark.to_spark()`
- No soporta streaming
  - Truco: se puede utilizar dentro de un `foreachbatch`

Ecosistema Flink



- Cambios actuales/futuros:
  - Table para datos estructurados y DataStream para (des)estructurados
  - PyFlink: alguna funcionalidad implementada



- Hadoop
- Ecosistema Hadoop v1
- Ecosistema Hadoop v2
- Ecosistema Spark
- Ecosistema Flink

Gracias

viu | Universidad Internacional de Valencia