Inteligencia de Negocio y Gestión de Procesos (INGP)

Seminario Big Data

Profesores:

Juan C. Trujillo, Alejandro Maté
LUCENTIA Research Group

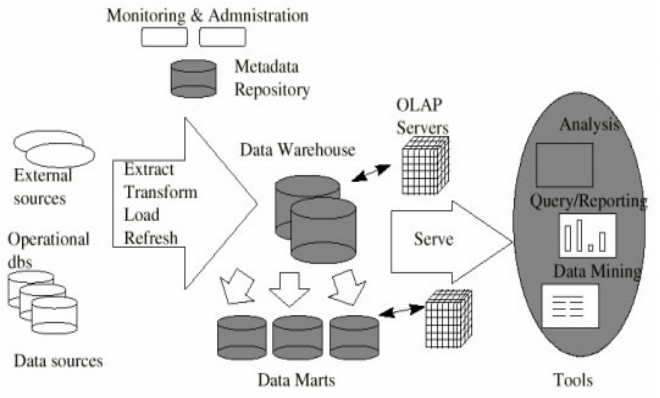
Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

1

Seminario de Big Data

- Arquitectura tradicional de BI

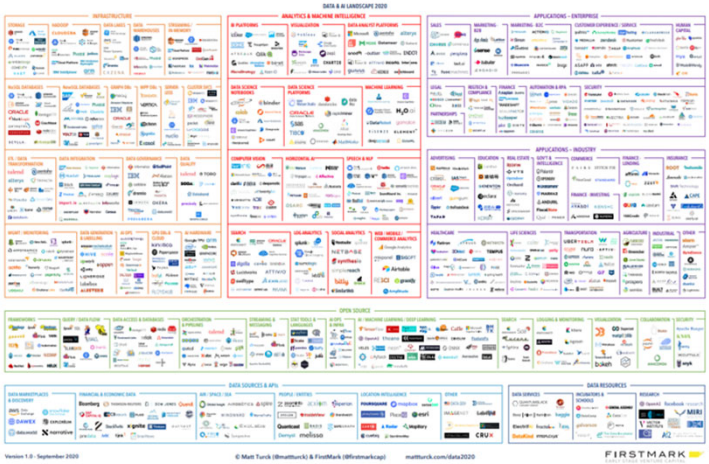


2

Seminario de Big Data

3

- Big Data Landscape 2020



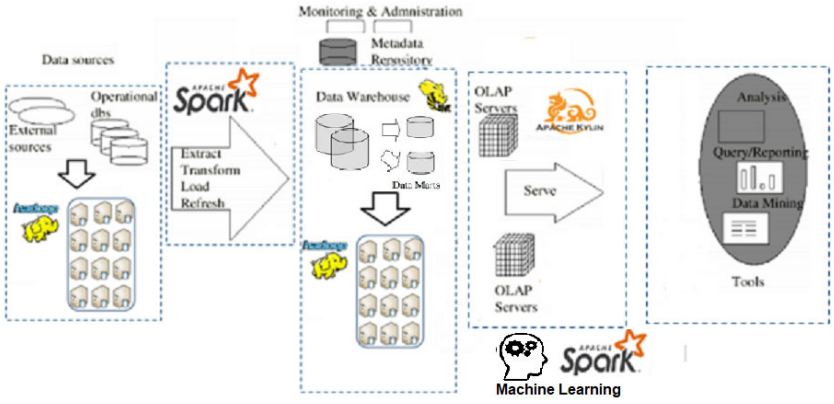
INGP. 2021

3

Seminario de Big Data

4

- Arquitectura BI con tecnología Big Data



INGP. 2021

4

Fuentes de Big Data

5

- Variabilidad de formatos de datos abiertos en 2017

	format	$\|resources\|$	%	$\|portals\|$
1	**HTML**	491,891	25	74
2	PDF	182,026	9.2	83
3	**CSV**	179,892	9.1	108
4	XLS(X)	120,703	6.1	89
5	**XML**	90,074	4.6	79
6	**ZIP**	50,116	2.5	74
	...			
11	**JSON**	28,923	1.5	77
16	**RDF**	10,445	0.5	28

INGP. 2021

5

Fuentes de Big Data

6

- Incluso tratándose del mismo formato podemos encontrar otros problemas:
  - Estructura y metadatos
    - Ej. El orden y el número de columnas no coinciden, nombres distintos, no se entiende que representan, etc
  - Accesibilidad
    - Ej. La información recopilada de fuentes externas puede dejar de ser accesible en cualquier momento
  - Confianza en la procedencia de datos
    - Confianza del proveedor: ¿cuánto confiamos en quien nos proporciona los datos?
  - Multilingüismo y semántica
    - Ej. La fecha se registra de manera diferente en inglés 01/30/2021 que en español 30/01/2021

INGP. 2021

6

Herramientas Big Data

7

- Para realizar análisis sobre fuentes de Big Data, necesitamos herramientas que faciliten nuestra tarea:
  - Flexibilidad para leer múltiples y variados formatos de datos
  - Altas capacidades para procesar, analizar y visualizar los datos
  - Capacidad para escalar el procesamiento de datos

7

Apache Spark

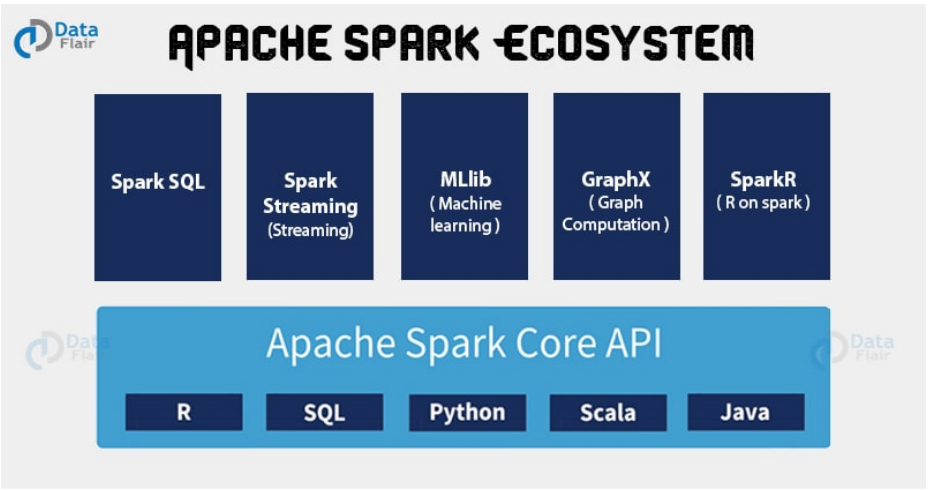
8

- Es un motor de procesamiento de datos distribuido.
- Se utiliza tanto para tareas de ciencia de datos como para procesamiento de datos a gran escala (TB y PB de datos).
- Creado en la universidad de Berkeley en California en 2009. Liberado como Open Source en 2010 con licencia BSD. En 2013 Spark fue donado a Apache Software Foundation
- Apache Spark es mantenido por la Comunidad con más de 45,903 commits y 1971 contribuyentes (a lo largo de su ciclo de vida).
- Se considera el primer software de código abierto que hace que la programación distribuida sea realmente accesible para científicos de datos. Se utiliza para procesar y analizar una gran cantidad de datos.

8

Apache Spark

9



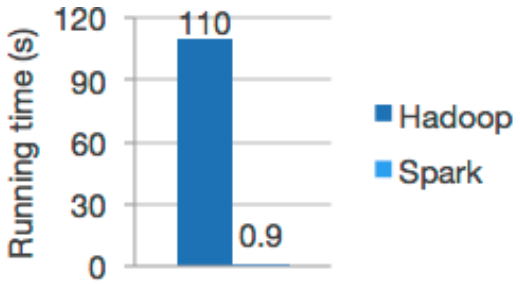
INGP. 2021

9

Apache Zeppelin

10

- Apache Spark logra un alto rendimiento tanto para datos por lotes como streaming mediante DAG scheduler, RDD (Resilient Distributed Datasets (tipo de dato basico), un optimizador de consultas y motor de ejecución física



INGP. 2021

10

Apache Zeppelin

11

- Apache Zeppelin es una aplicación de formato **notebook** que nos permite usar gran variedad de leguajes:
  - Apache Spark
  - Elastic Search
  - Python
  - R
  - SQL sentences
  - …
- Viene con su propio conjunto de visualizaciones (Helium)
- Y lo más importante, es de **código abierto**

INGP. 2021

11

Apache Zeppelin

12

- Similar a aplicaciones como Jupyter, Databricks

```
%pyspark
FireDF.registerTempTable('FireDFTable')
```

```
%sql
SELECT * from FireDFTable
```

Call Number	Unit ID	Incident Number	Call Type	Call Date	Watch Date	Received DtTm	Entry DtTm	Dispatch DtTm
1030101	E18	306091	Medical Incident	04/12/2000	04/12/2000	04/12/2000 09:00:29 PM	04/12/2000 09:01:40 PM	04/12/2000 09:02:00 PM
1030104	M14	30612	Medical Incident	04/12/2000	04/12/2000	04/12/2000 09:09:02 PM	04/12/2000 09:10:17 PM	04/12/2000 09:10:29 PM
1030106	M36	30614	Medical Incident	04/12/2000	04/12/2000	04/12/2000 09:09:44 PM	04/12/2000 09:10:56 PM	04/12/2000 09:11:47 PM
1030107	E01	30615	Alarms	04/12/2000	04/12/2000	04/12/2000 09:13:47 PM	04/12/2000 09:13:51 PM	04/12/2000 09:14:13 PM
1030108	RS1	30616	Medical Incident	04/12/2000	04/12/2000	04/12/2000 09:14:43 PM	04/12/2000 09:16:11 PM	04/12/2000 09:16:24 PM
1030112	T03	30620	Citizen Assist / Service Call	04/12/2000	04/12/2000	04/12/2000 09:24:27 PM	04/12/2000 09:24:54 PM	04/12/2000 09:25:10 PM
1030116	E38	30624	Electrical Hazard	04/12/2000	04/12/2000	04/12/2000 09:25:55 PM	04/12/2000 09:28:06 PM	04/12/2000 09:28:46 PM
1030117	E15	30626	Odor (Strange / Unknown)	04/12/2000	04/12/2000	04/12/2000 09:27:55 PM	04/12/2000 09:28:38 PM	04/12/2000 09:30:27 PM

Zeppelin

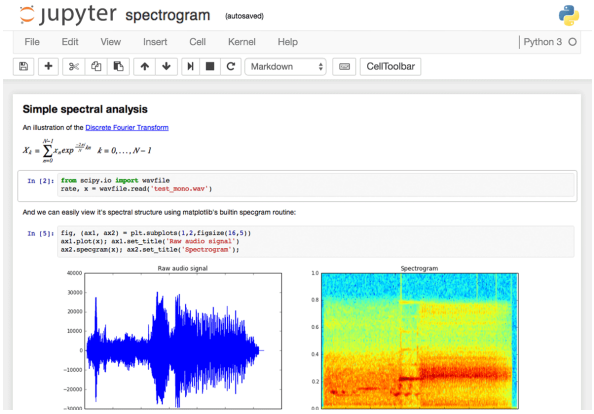
INGP. 2021

12

Apache Zeppelin

13

- Similar a aplicaciones como Jupyter o Databricks



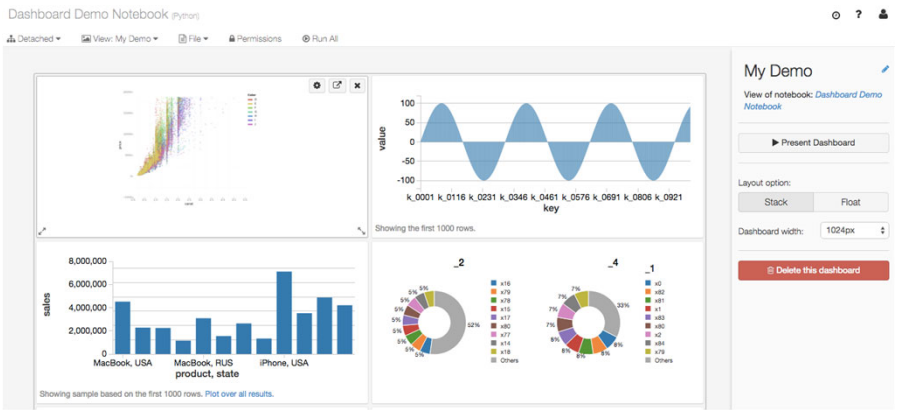
Jupyter

13

Apache Zeppelin

14

- Similar a aplicaciones como Jupyter o Databricks



Databricks

14

Apache Zeppelin

15

- Zeppelin se encarga de gestionar la **comunicación** entre el **notebook** y los diferentes **intérpretes**
- Nos permite usar **comandos especiales,** como z.show(datos) que permite visualizar los datos en diferentes visualizaciones
- Nos permite combinar **diferentes lenguajes** dependiendo del que mejor se adapte a nuestras necesidades en cada momento
  - Permite procesar datos con Spark, y luego consultarlos con SQL
- Podemos **publicar las visualizaciones creadas** en una página web
  - ¡Cuidado con los problemas de seguridad!

15

Apache Zeppelin

16

- Cargar un CSV:
  - df = spark.read.csv('[path to file]', inferSchema=True, header=True);
- Representar datos:
  - **df.show();**

```
# Looks the data of the DataFrame.
df.show()

# output
+--------------+------+
|     name     | age  |
+--------------+------+
| Juan carlos  |  54  |
| Alejandro    |  42  |
| Jose Manuel  |  37  |
| Alex         |  28  |
+--------------+------+
```

```
# operations on the data
# group by column "age"
# order descendent by column "age"
df
.groupBy("age")
.orderBy(desc("age"))
.count()
.show()

# output
+-----+-------+
| age | count |
+-------------+
| 54  |   1   |
| 42  |   1   |
| 37  |   1   |
| 28  |   1   |
+-------------+
```


16

Apache Zeppelin

17

- Cargar un CSV:
  - df = spark.read.csv('[path to file]', inferSchema=True, header=True);
- Representar datos:
  - **z.show(df);**

```
z.show(df)
df: org.apache.spark.sql.DataFrame = [first_column: int, second_column: int ... 1 more field]
```

first_column	second_column	third_column
1	1	1
2	2	2
3	3	3

INGP. 2021

17

¿Por dónde empezar?

18

- BIG DATA projects aren't one man thing
  - Servidores
  - Arquitectura
  - Programación
  - Diseño
  - Análisis
  - Dirección
    - DevOps, Backend, Frontend, Data scientist…

INGP. 2021

18

¿Por dónde empezar?

19

- Data scientist
  - …a data scientist is 1) a data analyst in California or 2) a statistician under 35
    - Gartner blog post by analyst Svetlana Sicular
    - Estadística
    - R, Matlab, SAS, SPSS
    - Minería de datos
    - Procesamiento de lenguaje natural
    - Machine Learning
    - Map/Reduce, Hadoop, Hive, etc
    - Python
  - The notion of a Data Scientist is a little mad but then so is Big Data. Removing the buzzwords just leaves you with….*Data*.

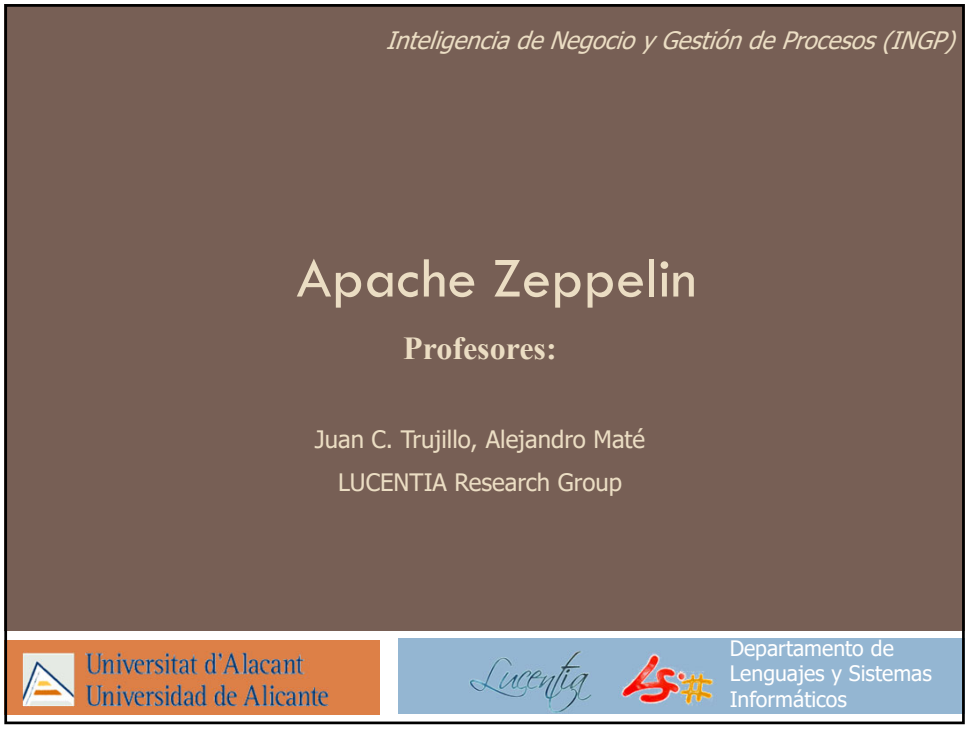
19

¿Por dónde empezar?

20

- ¿Por dónde empezar?
  - Realizar el trabajo de la asignatura con Spark y Zeppelin.
  - Cursos de big data
    - https://unimooc.com/curso-big-data/
    - https://cloud.google.com/certification/data-engineer
    - https://www.coursera.org/specializations/scala#courses
      https://www.coursera.org/learn/scala-spark-big-data#instructors
  - Cursos Data Science
    - https://web.ua.es/es/verano/2021/campus/introduccion-al-deep-learning.html
    - https://cloud.google.com/certification/machine-learning-engineer
  - (Película) El gran Hackeo
  - (Película) Sesgo codificado
  - (Película) El dilema de las redes

20


Inteligencia de Negocio y Gestión de Procesos (INGP)
Apache Zeppelin
Profesores:
Juan C. Trujillo, Alejandro Maté
LUCENTIA Research Group
Universitat d'Alacant
Universidad de Alicante
Lucentia
Departamento de
Lenguajes y Sistemas
Informáticos

21