01MBID Fundamentos de la tecnología Big Data

Universidad Internacional de Valencia

Sesión 4

Tema 3

De:
Planeta Formación y Universidades

Agenda

- **Dudas**
- **Seminario**
- **Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles**
- **Práctica MongoDB Atlas + Compass**

Agenda

- **Dudas**
- **Seminario**
- **Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles**
- **Práctica MongoDB Atlas + Compass**

Agenda

- Dudas
- **Seminario**
- **Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles**
- **Práctica MongoDB Atlas + Compass**

SEMINARIO - Ágora de Ciencia y Tecnología - Sala del curso

Transformando los datos en valor gracias a IoT & Big Data - Gustavo Martín

20h a 22h

https://eu.bbcollab.com/guest/f006b695dd4142a585e4c4b1b4a813c4

Test, AEC 5%, les avisaremos por el campus

Agenda

- Dudas
- Seminario
- **Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles**
- **Práctica MongoDB Atlas + Compass**

Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles

Introducción

En los entornos **Big Data**, el **progreso y la innovación ya no** se ven obstaculizados por la **capacidad de recopilar datos**, **sino** por la **capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento** de los datos recopilados **de manera oportuna y en una forma escalable**.

Introducción

La elección del tipo de técnicas de procesamiento y análisis de datos influirá decisivamente en el resultado.

- Potencia
- Escalabilidad
- Valores atípicos (outliers)
- Fraudes
- Seguridad
- Latencia

Estructuras de Datos y Tecnologías para Selección de Datos Útiles

1) Procesamiento Batch

2) Procesamiento de flujo y tiempo Real

3) Arquitectura Lambda

4) RDF

1) Procesamiento Batch

2) Procesamiento de flujo y tiempo Real

3) Arquitectura Lambda

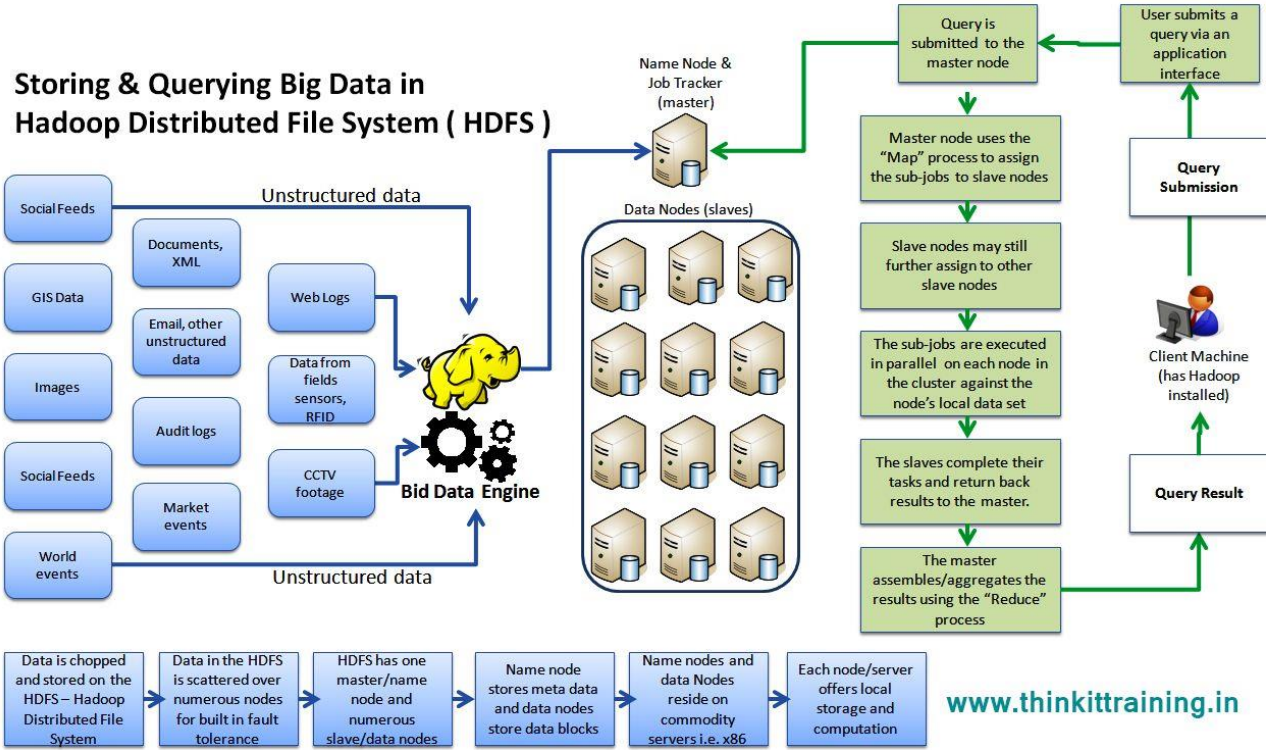
4) RDF

Procesamiento Batch

Procesamiento **por lotes** de grandes volúmenes de datos

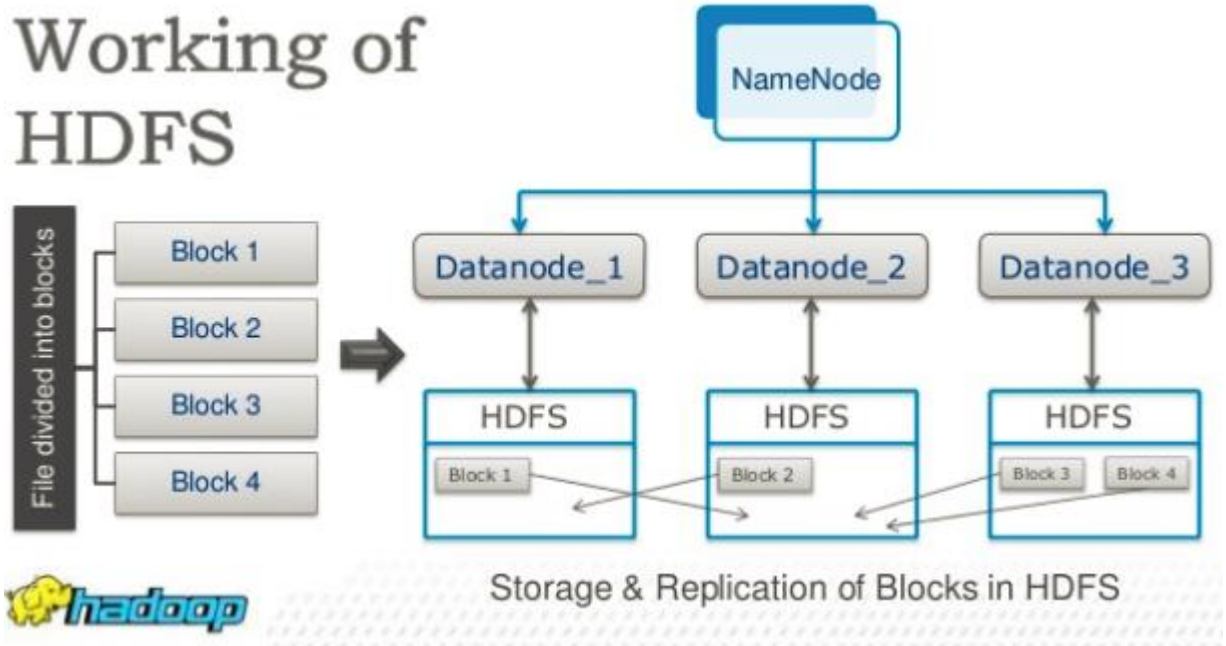
Herramienta más común: hadoop

Procesamiento Batch



Hadoop Distributed File System (HDFS)

- Gran Volumen
- Redundancia
- Escalabilidad
- Distribuido



Hadoop

Mejoras:

- Menores latencias.
- Minimización del tiempo de respuesta.
- Incremento de la Precisión.
- Soporte *stream*

Procesamiento Batch

Ejemplos:

- Cálculo del valor de mercado de los activos, que no necesita revisarse más de una vez al día.
- Cálculo mensual del coste de las facturas de teléfono de los empleados.
- Generación de informes relacionados con temas fiscales.

1) Procesamiento Batch

2) Procesamiento de flujo y tiempo Real

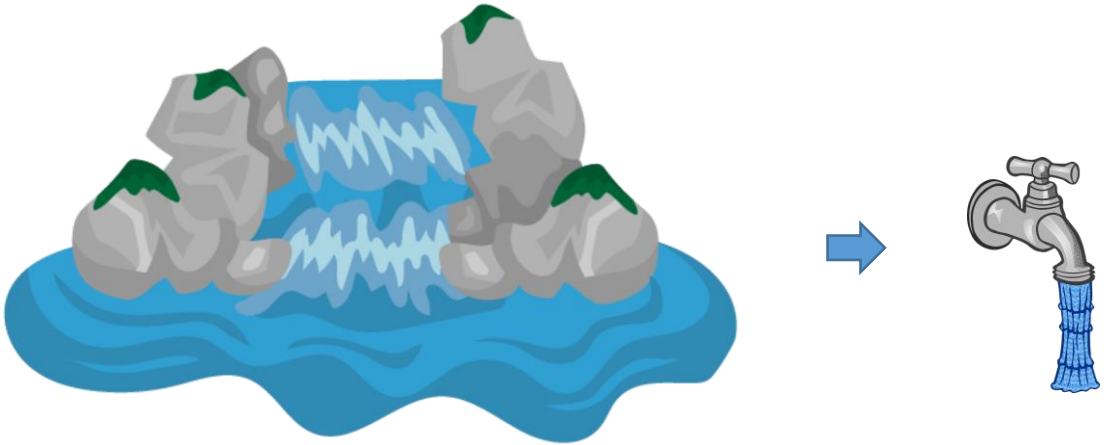
3) Arquitectura Lambda

4) RDF

Procesamiento de flujo y tiempo real

Modelo en el que los datos asociados a series de tiempo (hechos), se generan y procesan de forma continua y sin interrupción. (*Streaming*)

En cada momento procesamos una “pequeña” cantidad de datos.



Procesamiento de flujo y tiempo real

	Flujo *	Tiempo Real
Datos procesados	% aceptable	100%
Tiempo procesamiento	Razonable	En vivo

*Dependerá de la aplicación y contexto particular

Procesamiento de flujo y tiempo real

Limitaciones:

- Memoria suficiente para almacenar entradas en cola.
- Velocidad de procesamiento a largo plazo más rápida, o por lo menos igual, a la velocidad de entrada de datos. O tendremos desbordamiento de datos.
- La mayoría de técnicas sólo pueden gestionar un subconjunto de datos.
- En tiempo real es difícil garantizar que no se pierden datos.
- Tiempo real (100%) es extremadamente complejo y costoso.

Procesamiento de flujo y tiempo real

Herramientas *Full streaming*:

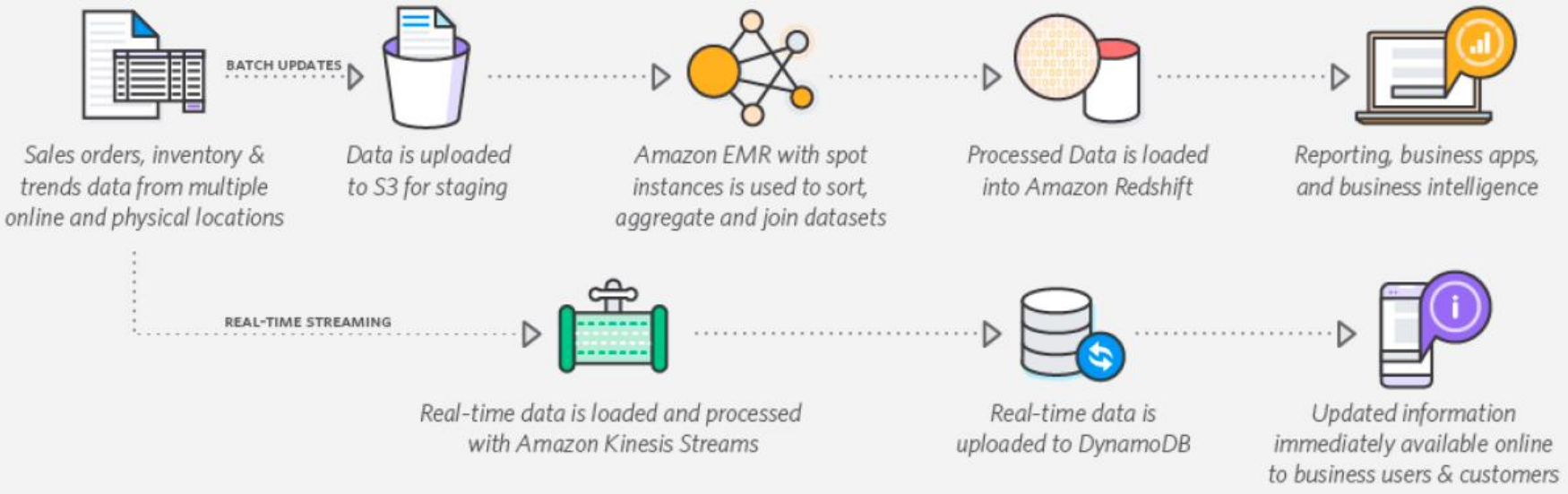
- Apache Storm
- Apache Samza
- Apache Flink
- AWS Kinesis

Herramientas *Microbatch*:

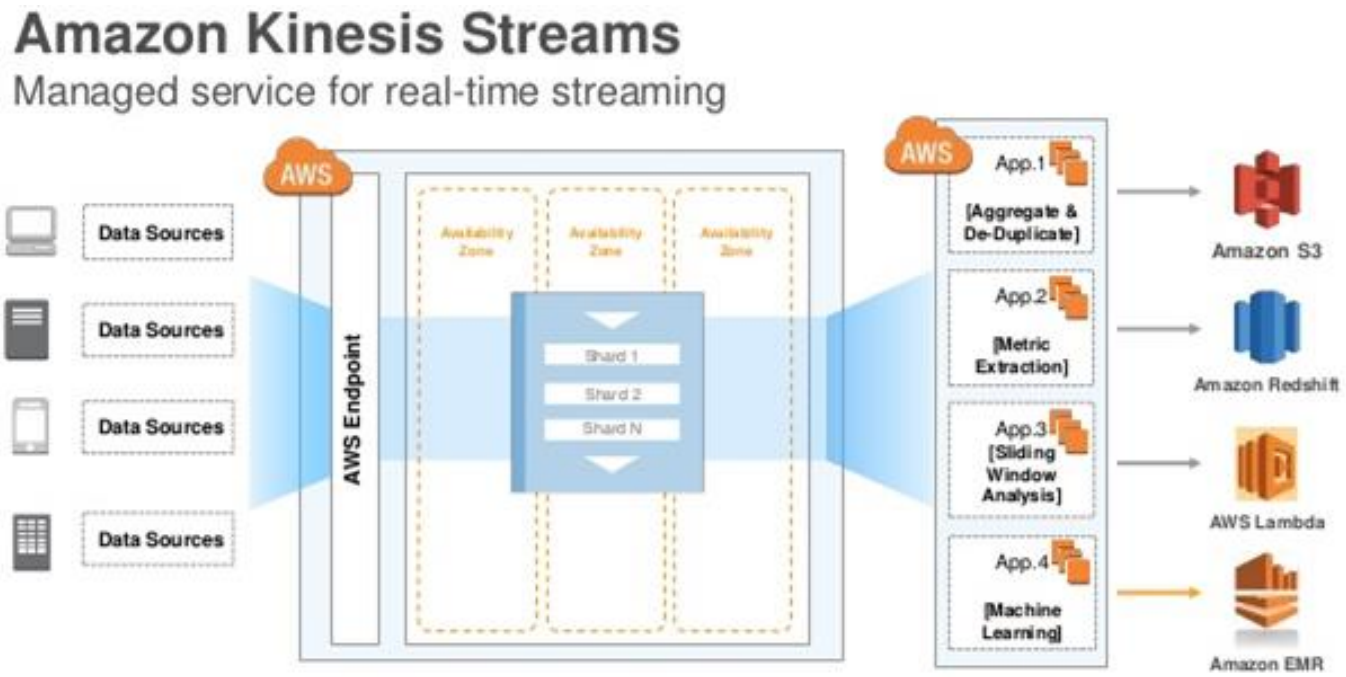
- Spark Streaming
- Storm Trident

Procesamiento de flujo y tiempo real

Procesamiento de flujo y tiempo real - AWS



Procesamiento de flujo y tiempo real - AWS



Más herramientas

Flume: ingesta de datos en entornos casi tiempo real [near real-time (NRT)]
https://flume.apache.org/

Kafka: sistema de almacenamiento distribuido y replicado.
http://kafka.apache.org/

Hive: infraestructura de data-warehose construida sobre Hadoop
https://hive.apache.org/

RabbitMQ: sistema de almacenamiento distribuido para gestión de colas
https://www.rabbitmq.com/

LogStash: procesamiento de logs
http://logging.apache.org/

Loggly: monitoreo de logs
https://www.loggly.com/

1) Procesamiento Batch

2) Procesamiento de flujo y tiempo Real

3) Arquitectura Lambda

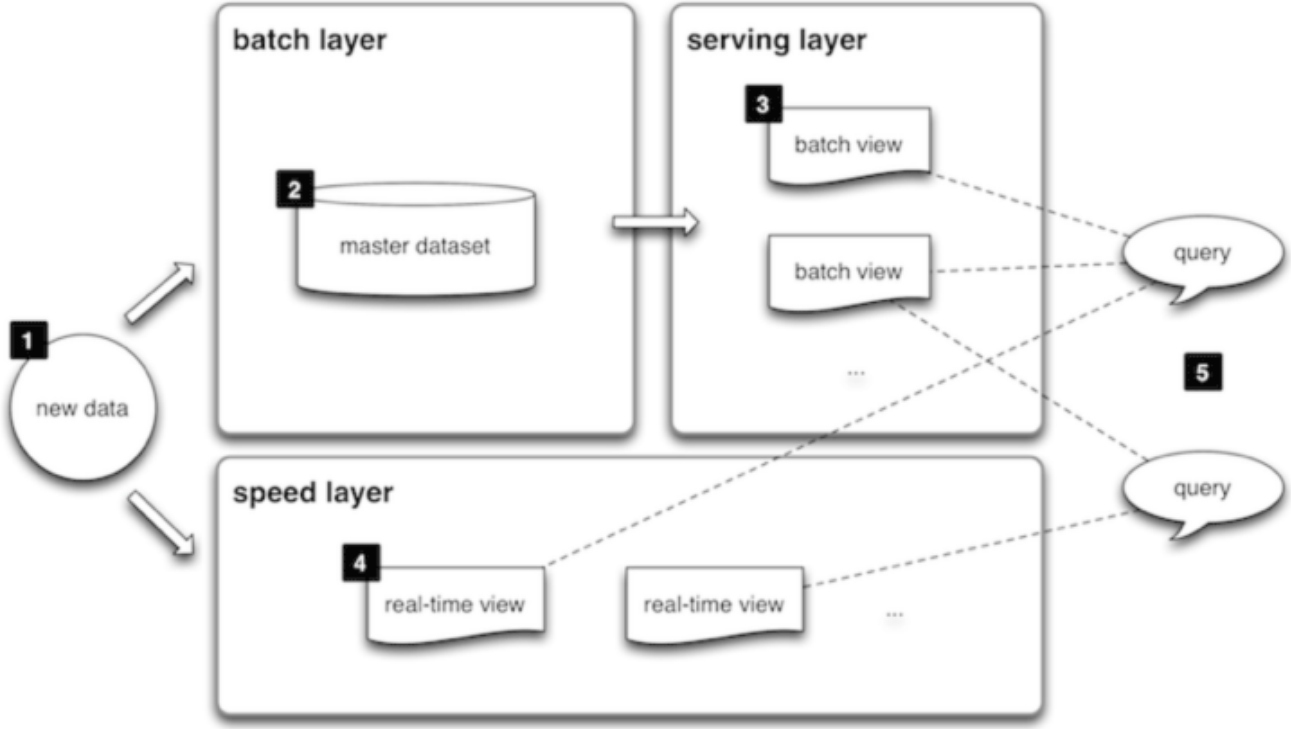
4) RDF

Arquitectura Lambda

Combina parte de solución Batch y parte de solución en Tiempo Real.

Arquitectura de procesamiento de datos genérica, escalable y tolerante a fallos.

Arquitectura Lambda



** Hausenblas & Bijnens (2017).*

Arquitectura Lambda

1) Todos los datos se envían a la capa de **procesamiento por lotes y** a la **capa de velocidad**.
2) La **capa** de **procesamiento por lotes** tiene dos funciones:
   - Administrar el conjunto de **datos maestro**.
   - Calcular previamente las **vistas por lotes**.
3) La **capa de servicio indexa las vistas** de lote para que puedan ser consultadas de manera de baja latencia y ad-hoc.
4) La **capa de velocidad** compensa la alta latencia de las actualizaciones en la capa de servicio y trata **sólo con datos recientes**.
5) **Las consultas entrantes** se pueden responder fusionando los resultados de las **vistas por lotes** y las **vistas en tiempo real**.

Arquitectura Lambda

Un caso de uso real para una arquitectura Lambda podría ser un sistema que recomiende libros en función de los gustos de los usuarios. Por un lado, tendría una capa batch encargada de para entrenar el modelo e ir mejorando las predicciones; y por otro, una capa streaming capaz de encargarse de las valoraciones en tiempo real.

Arquitectura Lambda – Para profundizar

Lambda architecture for cost-effective batch and speed big data processing

https://www.researchgate.net/publication/308871780_Lambda_architecture_for_cost-effective_batch_and_speed_big_data_processing

Lambda Architecture – AWS

https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

Ellis, Byron, and Justin Langseth. ***Real-Time Analytics : Techniques to Analyze and Visualize Streaming Data***, John Wiley & Sons, Incorporated, 2014. *ProQuest Ebook Central*, https://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1719571

1) Procesamiento Batch

2) Procesamiento de flujo y tiempo Real

3) Arquitectura Lambda

4) RDF

RDF

***R**esource **D**escription **F**ramework (**RDF**)*

- Modelo de datos para metadatos.
- Es un grafo dirigido y etiquetado de formato de datos para representar la información en la web.

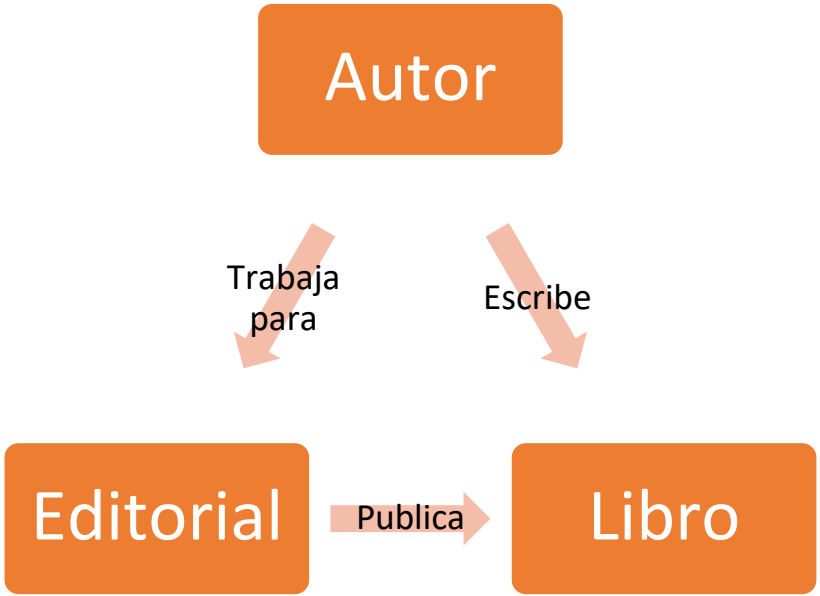
RDF

***S**PARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage*

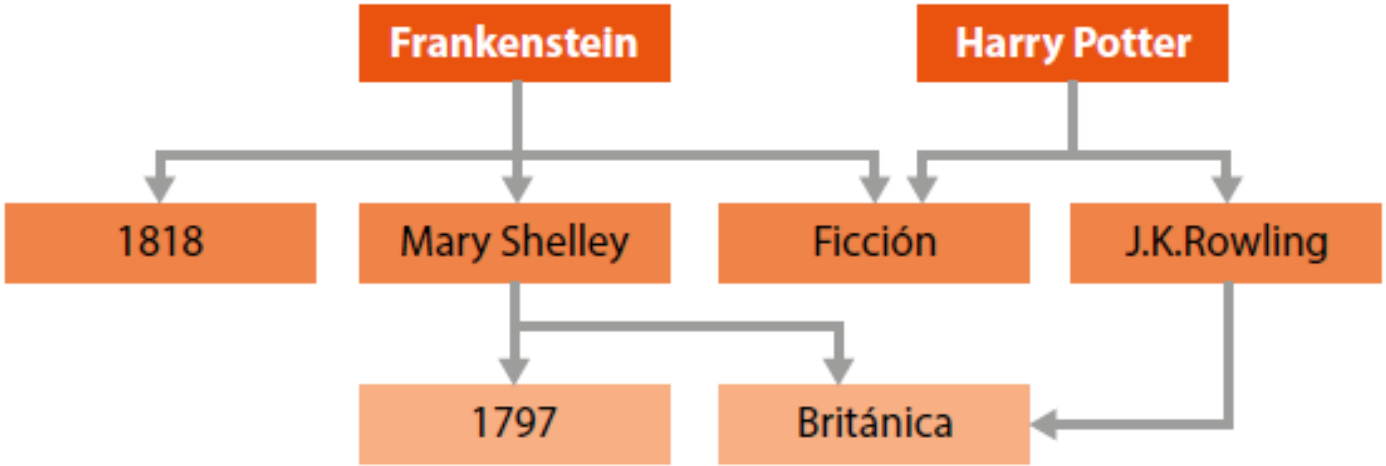
Lenguaje estandarizado para la consulta de grafos RDF.

https://www.w3.org/TR/rdf-sparql-query/

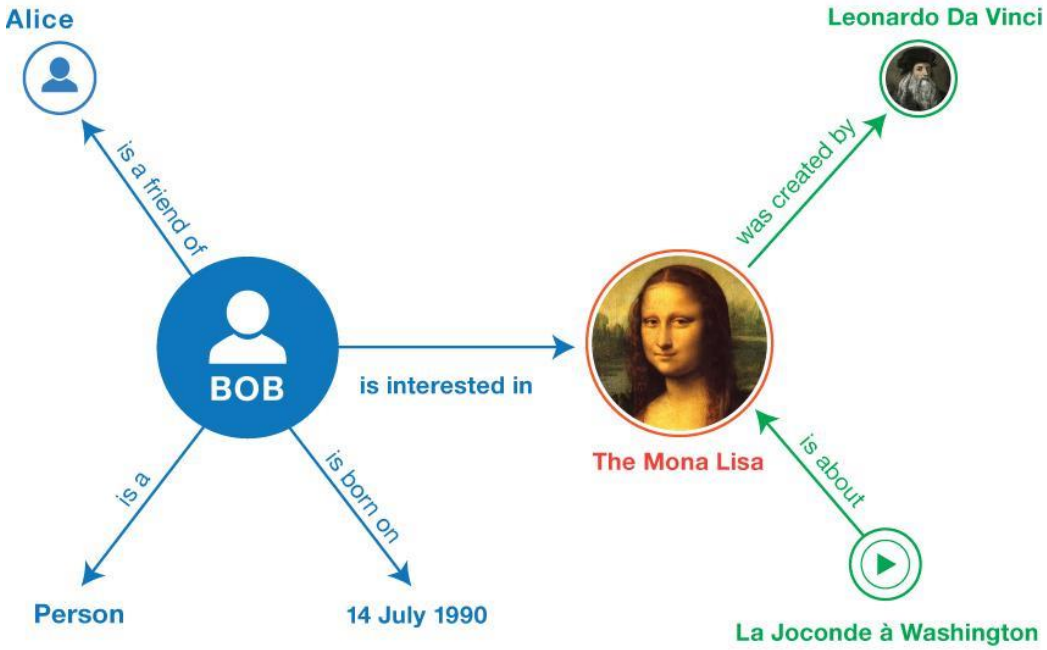
RDF



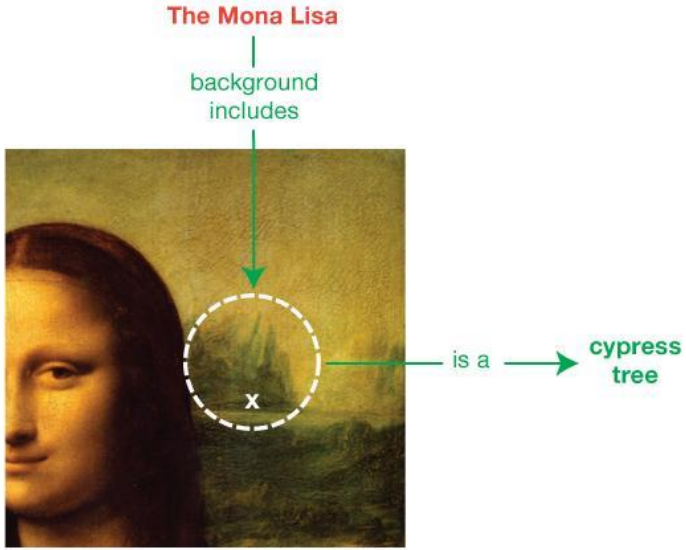
RDF


Frankenstein
Harry Potter
1818
Mary Shelley
Ficción
J.K.Rowling
1797
Británica

RDF


Alice
is a friend of
BOB
is interested in
is a
is born on
Person
14 July 1990
Leonardo Da Vinci
was created by
The Mona Lisa
is about
La Joconde à Washington

RDF



RDF

• Base de Datos RDF: BD orientada a grafos, estructuran la BD como un grafo o red. El conjunto de recursos o nodos están conectados entre sí mediante aristas (o enlaces) que describen las relaciones establecidas entre dichos recursos y/o nodos.

• Linked Open Data (LOD): Las BD RDF admiten el uso de URL, estas pueden estar accesibles online y también pueden enlazarse a otras bases de datos.

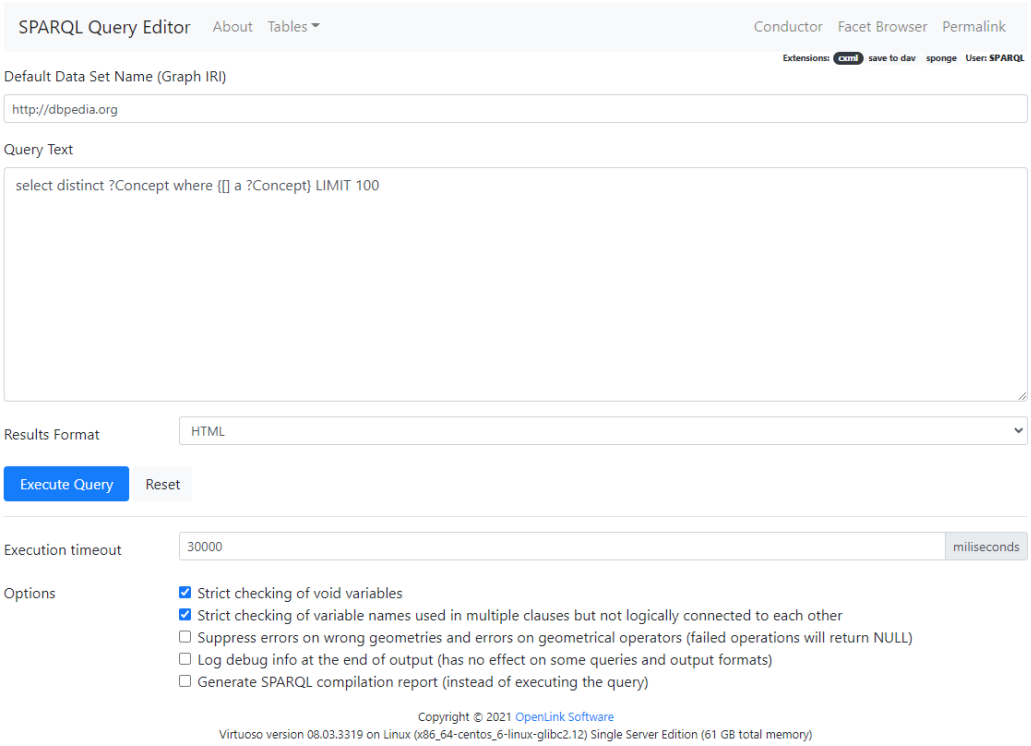
RDF

SPARQL es el lenguaje utilizado para interrogar este tipo de bases de datos

https://es.dbpedia.org/

Universidad Internacional de Valencia

https://dbpedia.org/sparql



Ejemplos de consultas

¿Actores/Actrices casados con Actores/Actrices?

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT ?actor ?actor2 WHERE{
  ?actor rdf:type dbpedia-owl:Actor .
  ?actor2 rdf:type dbpedia-owl:Actor .
  ?actor dbpedia-owl:spouse ?actor2 .
}
```

Ejemplos de consultas

PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

Qué ontología utilizaré para definir los términos:
-Actores/Actrices son tipo ‘Actor’
-La relación entre los nodos de casados es ‘spouse’

SELECT ?actor ?actor2 WHERE{

Seleccionar (SELECT) actor y actor2 (Variables) cuando se cumplan las condiciones (WHERE)

Ejemplos de consultas

?actor rdf:type dbpedia-owl:Actor .

actor debe ser del tipo Actor

?actor2 rdf:type dbpedia-owl:Actor .

actor2 debe ser del tipo Actor

?actor dbpedia-owl:spouse ?actor2 .

Debe existir una relación de spouse entre actor y actor2

Ejemplos de consultas

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT ?actor ?actor2 WHERE{
  ?actor rdf:type dbpedia-owl:Actor .
  ?actor2 rdf:type dbpedia-owl:Actor .
  ?actor dbpedia-owl:spouse ?actor2 .
}
```

Sponging: Use only local data (including data retrieved before), but do not retrieve more

Results Format: HTML

Execution timeout: 0 milliseconds *(values less than 1000 are ignored)*

Options:
- [x] Strict checking of void variables
- [ ] Log debug info at the end of output (has no effect on some queries and output formats)
- [ ] Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see details)

Run Query | Reset

Ejemplos de consultas

Primeros resultados de la consulta en formato HTML, los links se pueden seguir para saber los datos completos de cada actor (nodo del grafo)

actor	actor2
http://es.dbpedia.org/resource/Kajol	http://es.dbpedia.org/resource/Ajay_Devgan
http://es.dbpedia.org/resource/Jerome_Courtland	http://es.dbpedia.org/resource/Polly_Bergen
http://es.dbpedia.org/resource/Alexandra_Bastedo	http://es.dbpedia.org/resource/Patrick_Garland
http://es.dbpedia.org/resource/Portland_Hoffa	http://es.dbpedia.org/resource/Fred_Allen
http://es.dbpedia.org/resource/Polly_Bergen	http://es.dbpedia.org/resource/Jerome_Courtland
http://es.dbpedia.org/resource/Fred_Allen	http://es.dbpedia.org/resource/Portland_Hoffa
http://es.dbpedia.org/resource/Patrick_Garland	http://es.dbpedia.org/resource/Alexandra_Bastedo
http://es.dbpedia.org/resource/George_Loane_Tucker	http://es.dbpedia.org/resource/Elisabeth_Risdon
http://es.dbpedia.org/resource/Maurizio_D'Ancora	http://es.dbpedia.org/resource/Sandra_Ravel
http://es.dbpedia.org/resource/Esmeralda_Moya	http://es.dbpedia.org/resource/Carlos_García_Cortázar
http://es.dbpedia.org/resource/Jane_Frazee	http://es.dbpedia.org/resource/Glenn_Tryon

Ejemplos de consultas

Por ejemplo si clicamos en la primera actriz ‘Kajol’ obtenemos:

About: Kajol

An Entity of Type : actor, from Named Graph : http://es.dbpedia.org, within Data Space : es.dbpedia.org

Kajol Devgan, de soltera Kajol Mukherjee (en idioma hindi, काजोल; Bombay, 5 de agosto de 1974) es una actriz de cine india. Descendiente Bengalí-Marathi, pertenece a una de las tres familias más respetadas en el negocio de la cinematografía, la familia Mukherjee-Samarth. Está casada con el actor indio Ajay Devgan con el que tiene dos hijos.

Property	Value
dbo:abstract	Kajol Devgan, de soltera Kajol Mukherjee (en idioma hindi, काजोल; Bombay, 5 de agosto de 1974) es una actriz de cine india. Descendiente Bengalí-Marathi, pertenece a una de las tres familias más respetadas en el negocio de la cinematografía, la familia Mukherjee-Samarth. Está casada con el actor indio Ajay Devgan con el que tiene dos hijos. (es)
dbo:birthName	Kajol Mukherjee (es)
dbo:birthPlace	dbpedia-es:Bombay dbpedia-es:India
dbo:occupation	dbpedia-es:Actriz
dbo:spouse	dbpedia-es:Ajay_Devgan

rdf:type

- owl:Thing
- foaf:Person
- dbo:Person
- dbo:Actor
- dul:NaturalPerson
- schema:Person
- dul:Agent
- wikidata:Q215627
- wikidata:Q24229398
- wikidata:Q33999
- wikidata:Q483501
- wikidata:Q5
- dbo:Agent
- dbo:Artist

Ejemplos de consultas

Ya sea en el resultado de la consulta o en la ficha de la actriz ‘Kajol’, podemos acceder a los datos de su esposo y obtenemos:

About: Ajay Devgan

An Entity of Type : actor, from Named Graph : http://es.dbpedia.org, within Data Space : es.dbpedia.org

Ajay Devgan (en hindi, विशाल देवगन, en punyabi: ਵਿਸ਼ਾਲ ਦੇਵਗਨ, Nueva Delhi, India, 2 de abril de 1969) es un actor del cine de la India. Ha ganado dos Premios de National Film y en 2008 dirigió la película con su esposa la actriz Kajol.

rdf:type	owl:Thing foaf:Person dbo:Person dbo:Actor	is dbo:spouse of	dbpedia-es:Kajol

Ejemplos de consultas

¿Países con el euro como moneda?

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?pais WHERE{
  ?pais rdf:type dbpedia-owl:Country .
  ?pais dbpedia-owl:currency ?moneda .

FILTER(?moneda = <http://es.dbpedia.org/resource/Euro>)

}
```

Ejemplos de consultas

?pais rdf:type dbpedia-owl:Country .

Que sea un nodo de tipo país

?pais dbpedia-owl:currency ?moneda .

Obtenemos el atributo currency (moneda) del país

FILTER(?moneda = <http://es.dbpedia.org/resource/Euro>)

Filtramos que esa moneda debe ser Euro

Ejemplos de consultas

SELECT DISTINCT ?pais WHERE{

Es necesario incluir DISTINCT para evitar duplicados. Por ejemplo España sin el distinct saldría 4 veces, puede ser error de la BD.

pais
http://es.dbpedia.org/resource/Reino_Gay_y_Lésbico_de_las_Islas_del_Mar_del_Coral
http://es.dbpedia.org/resource/Alemania
http://es.dbpedia.org/resource/Montenegro
http://es.dbpedia.org/resource/Principado_de_Seborga
http://es.dbpedia.org/resource/Chipre
http://es.dbpedia.org/resource/Kosovo
http://es.dbpedia.org/resource/Portugal

Ejemplos de consultas

Tenemos un país autoproclamado que también usa el euro. Siempre las consultas son sobre los datos que tengamos introducidos en la BD, en este caso en la DBpedia. Dbpedia obtiene los datos de Wikipedia.

El **Reino Gay y Lésbico de las Islas del Mar del Coral** (en inglés: *Gay and Lesbian Kingdom of the Coral Sea Islands*) fue una autoproclamada micronación no reconocida, establecida como una protesta política simbólica realizada por un grupo de activistas LGBT del sureste de Queensland, Australia. Es una expresión del nacionalismo queer.

https://es.wikipedia.org/wiki/Reino_Gay_y_L%C3%A9sbico_de_las_Islas_del_Mar_del_Coral

Ejemplos de consultas

Otros ejemplos de consultas, no está 100% actualizado, alguna no funcionan: https://es.dbpedia.org/wiki/Wiki.jsp?page=Ejemplos%20de%20consultas%20SPARQL#top

URL de interés

https://www.w3.org/standards/semanticweb/

https://wiki.dbpedia.org/

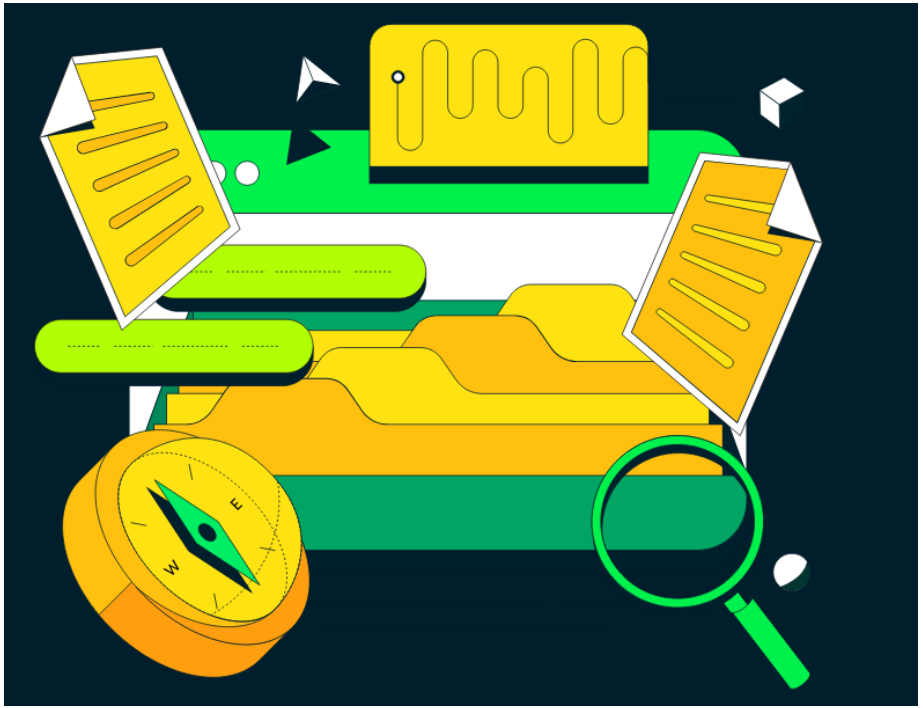
https://jena.apache.org/index.html

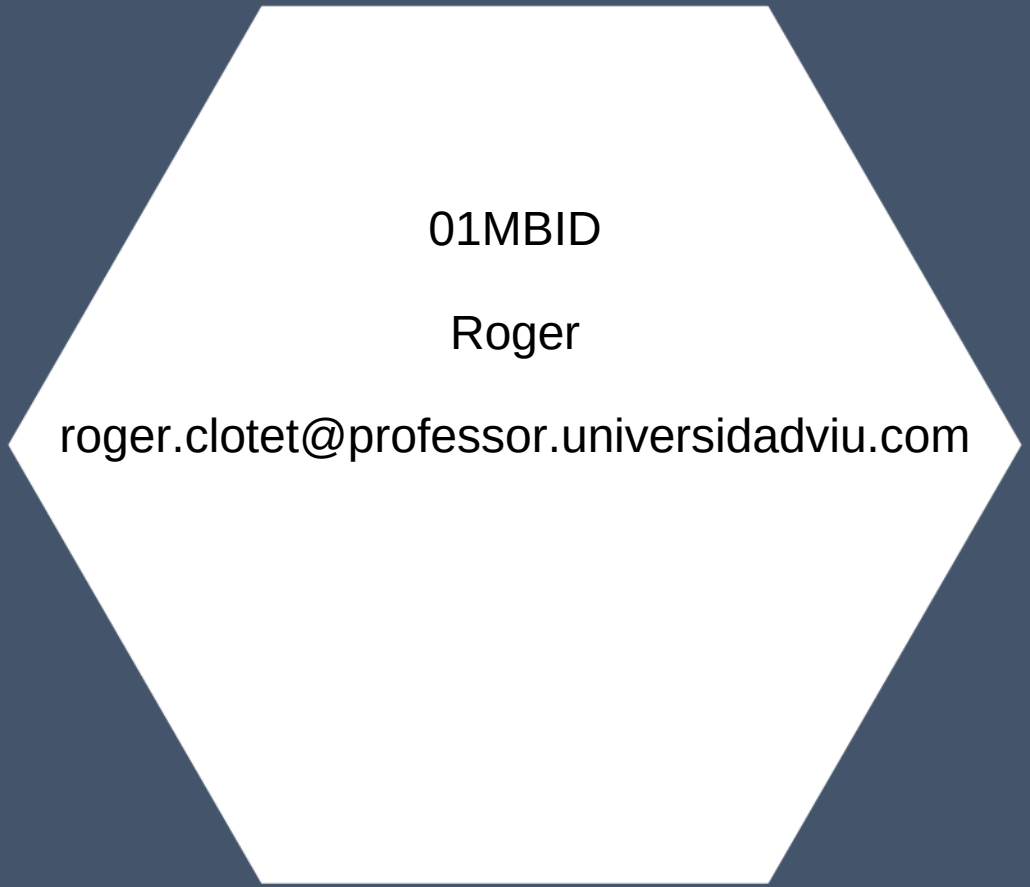
https://jena.apache.org/tutorials/sparql.html

Agenda

- Dudas
- Seminario
- Tema 3 - Estructuras de datos y tecnologías para selección de datos útiles
- **Práctica MongoDB Atlas + Compass**

Práctica MongoDB Atlas + Compass




01MBID
Roger
roger.clotet@professor.universidadviu.com

Gracias

Universidad
Internacional
de Valencia