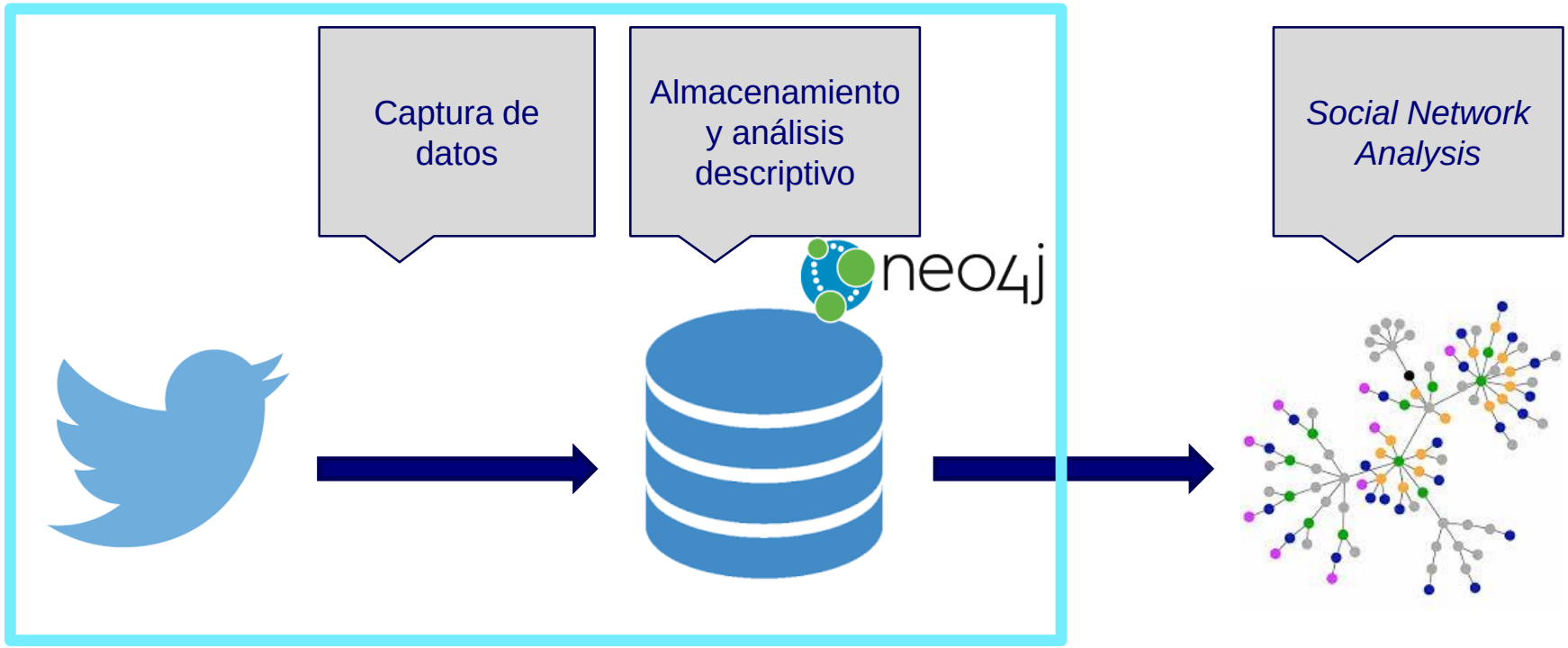
Uso de *Social Network Analytics* (SNA)

Estudis d'Informàtica, Multimèdia i Telecomunicació

Jordi Conesa i Caralt
14 de Junio de 2017

00.1 Vamos a ver *Social Network Analytics*?

Pues más bien poco…



00.1 Índice

- **¿Qué problema queremos resolver?**
- ¿Qué herramientas utilizaremos para hacerlo?
    - ¿Qué es una base de datos NoSQL?
    - ¿Qué es una base de datos en Grafo?
    - ¿Qué es un grafo?
    - ¿Qué es Neo4j?
- ¿Cómo se ha realizado la captura y almacenaje de datos?
- ¿Cómo se ha realizado el análisis de datos?

00.1 Análisis de datos de Twitter

Una empresa que ofrece un cliente de Twitter quiere analizar como utilizan Twitter sus usuarios para crear una nueva versión de su software.

00.1 Análisis de datos de Twitter

Se decide analizar la actividad en Twitter de 500 usuarios pre-seleccionados para obtener información sobre:

- **La actividad en la red**: media de tweets por usuario, proporción de retweets, proporción de tweets geolocalizados, patrones semanales de actividad, etc.
- **Aplicaciones que interactúan con Twitter**: que aplicaciones utilizan los usuarios, ranking de aplicaciones, etc.
- **Detección de usuarios importantes en la red**: usuarios con más seguidores, usuarios más bien conectados, usuarios más activos, etc.
- **Segmentación de comportamiento**: por temas de interés, por franjas horarias en qué se conectan, en función de sus seguidores, etc.

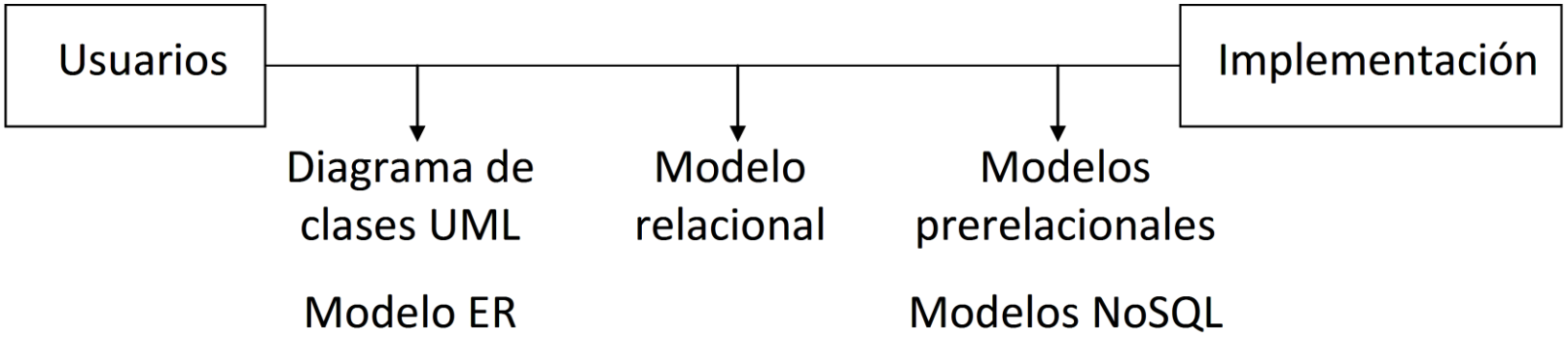
> 9.400.000 Registros

00.1 Índice

- ¿Qué problema queremos resolver?
- **¿Qué herramientas utilizaremos para hacerlo?**
    - ¿Qué es una base de datos NoSQL?
    - ¿Qué es una base de datos en Grafo?
    - ¿Qué es un grafo?
    - ¿Qué es Neo4j?
- ¿Cómo se ha realizado la captura y almacenaje de datos?
- ¿Cómo se ha realizado el análisis de datos?

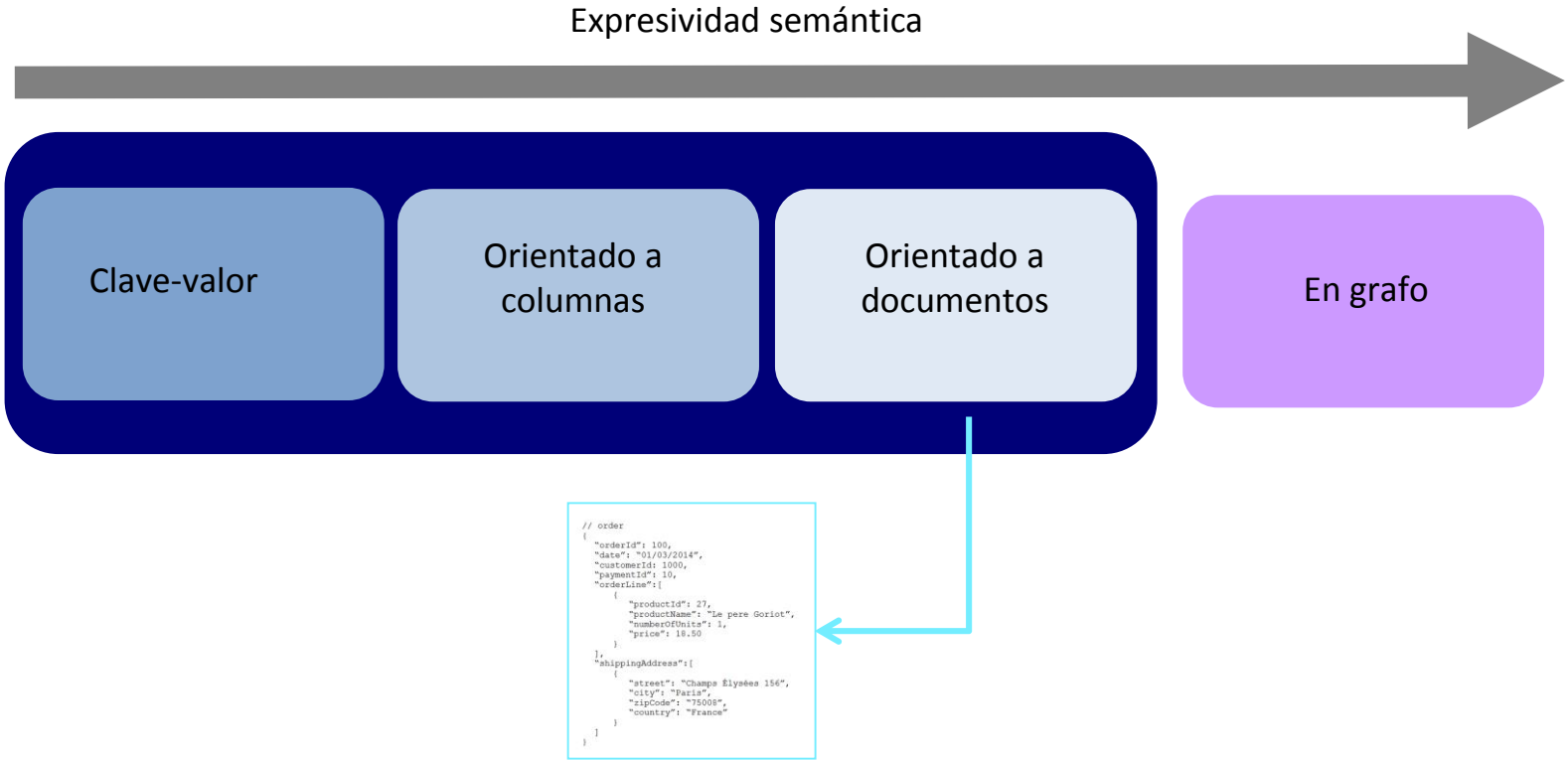
00.1 NoSQL

Término que aúna distintas familias de bases de datos que usan un **modelo de datos distinto** al relacional.

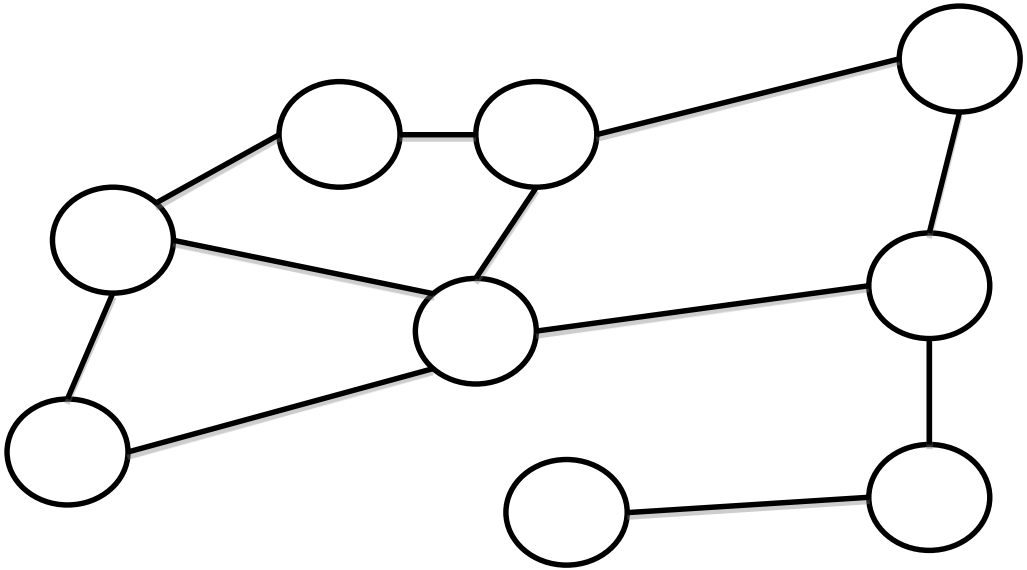


Conjunto de componentes que proporciona el sistema gestor de la base de datos para estructurar y manipular los datos

00.1 Familias de productos NoSQL



00.1 ¿Qué es un grafo?



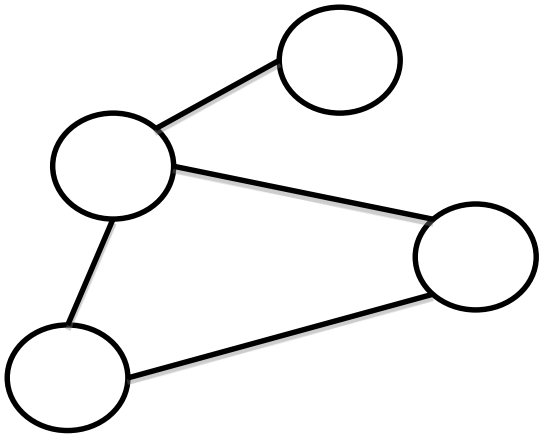
Nodos

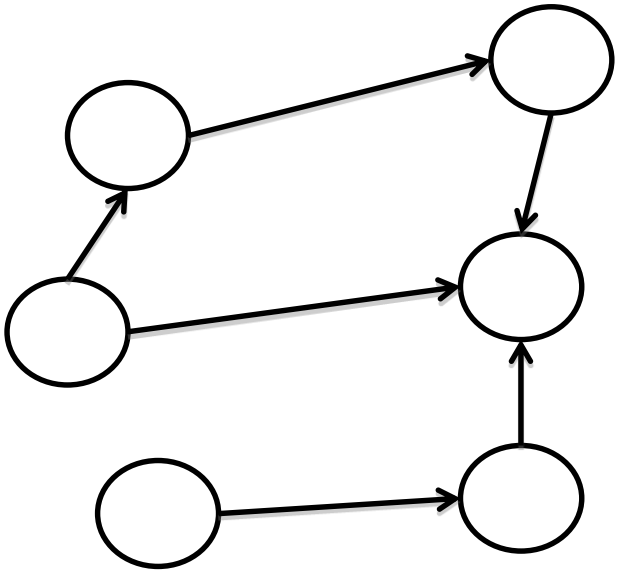
Aristas

00.1 Tipos de grafos

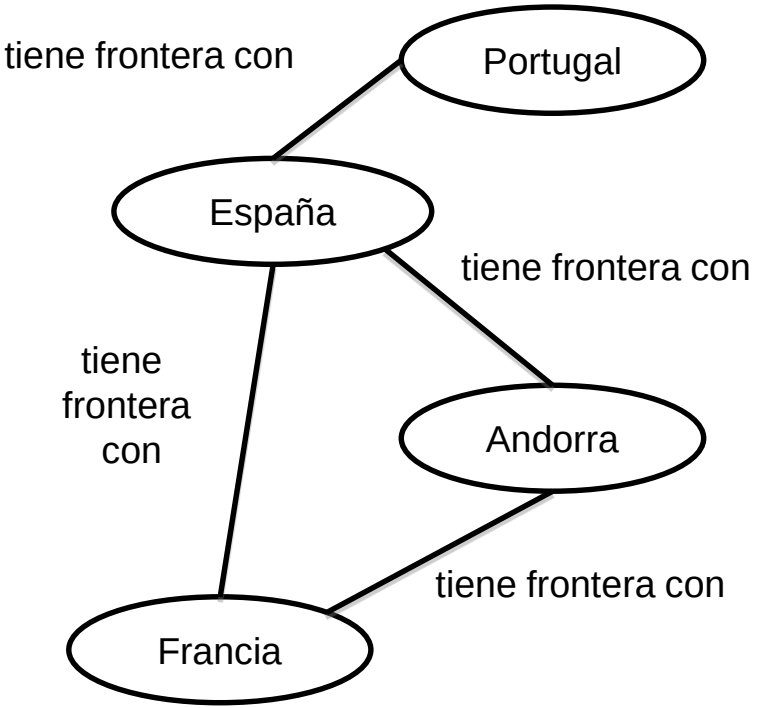
Tipos de aristas

Tipos de grafo

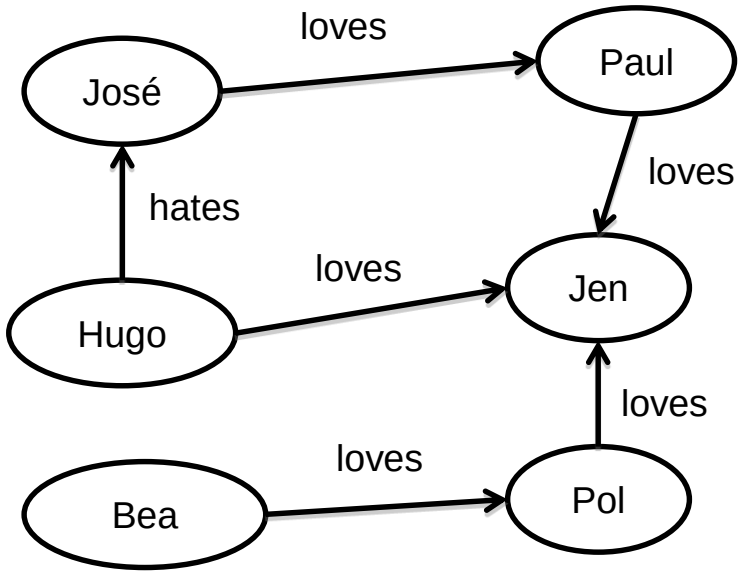




00.1 Uso de etiquetas en grafos



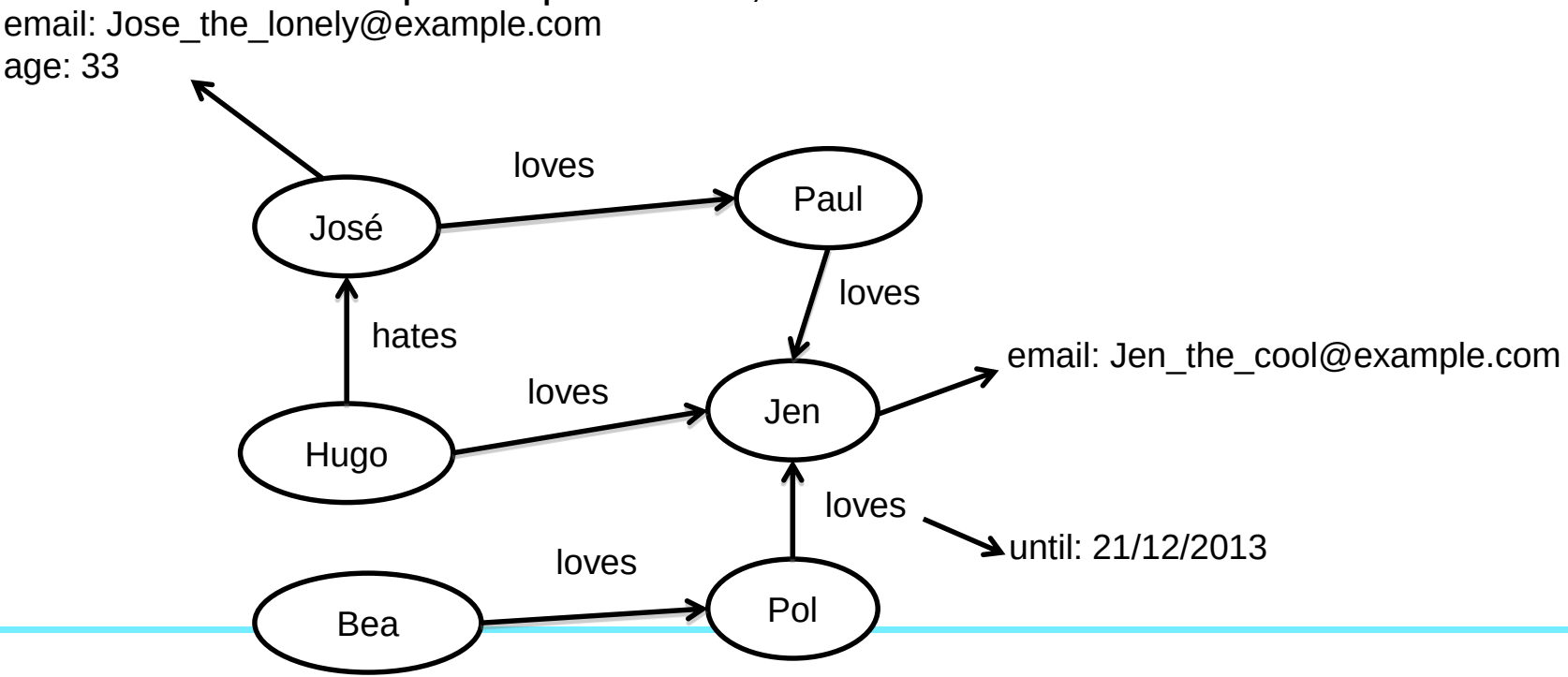
Grafo no dirigido etiquetado



Grafo dirigido etiquetado

00.1 Uso de propiedades en grafos

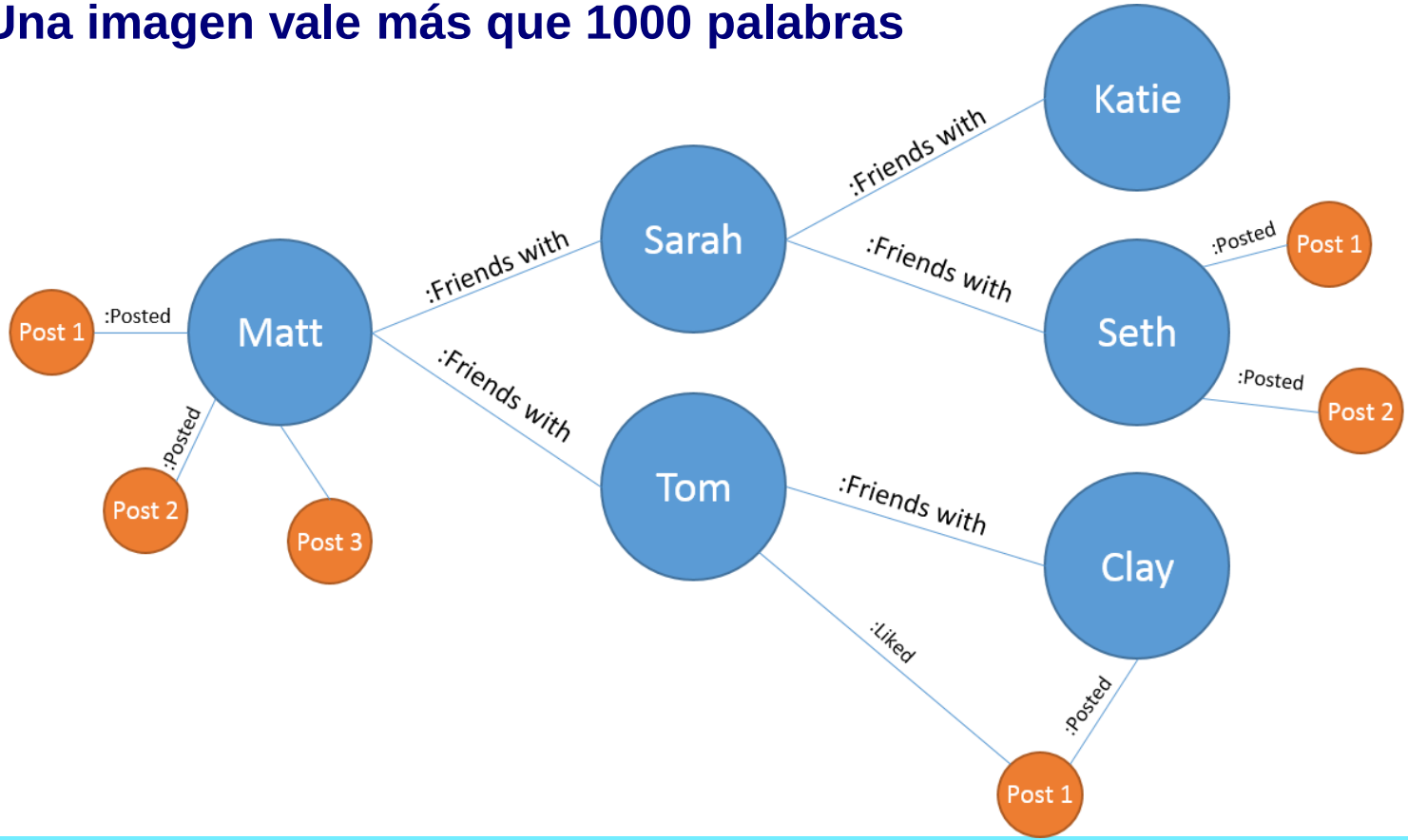
- Pueden asignarse tanto a nodos como a aristas.
- Están formadas por un par <clave, valor>.



00.1 ¿Qué es Neo4j?

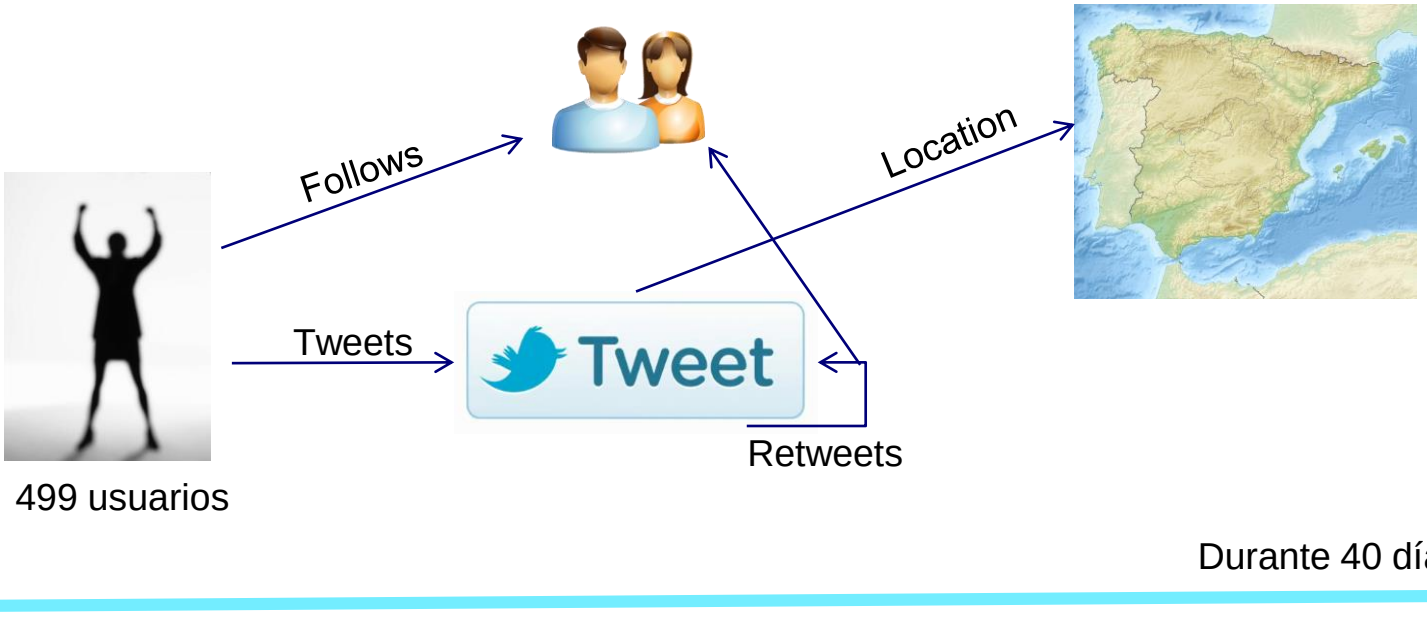
- Una base de datos NoSQL en grafo
- Sigue un modelo de grafo de propiedades etiquetado
- Disponible en sistemas Linux, Windows y OS X
- Accesible desde multitud de lenguajes de programación y de manipulación de grafos (gremlin, Cypher…)
- Usa un sistema de transacciones ACID
- Proporciona algunos algoritmos de grafos ya implementados
- Es la base de datos en grafo más popular hoy en día según db-engines.com

00.1 Una imagen vale más que 1000 palabras



00.1 Índice

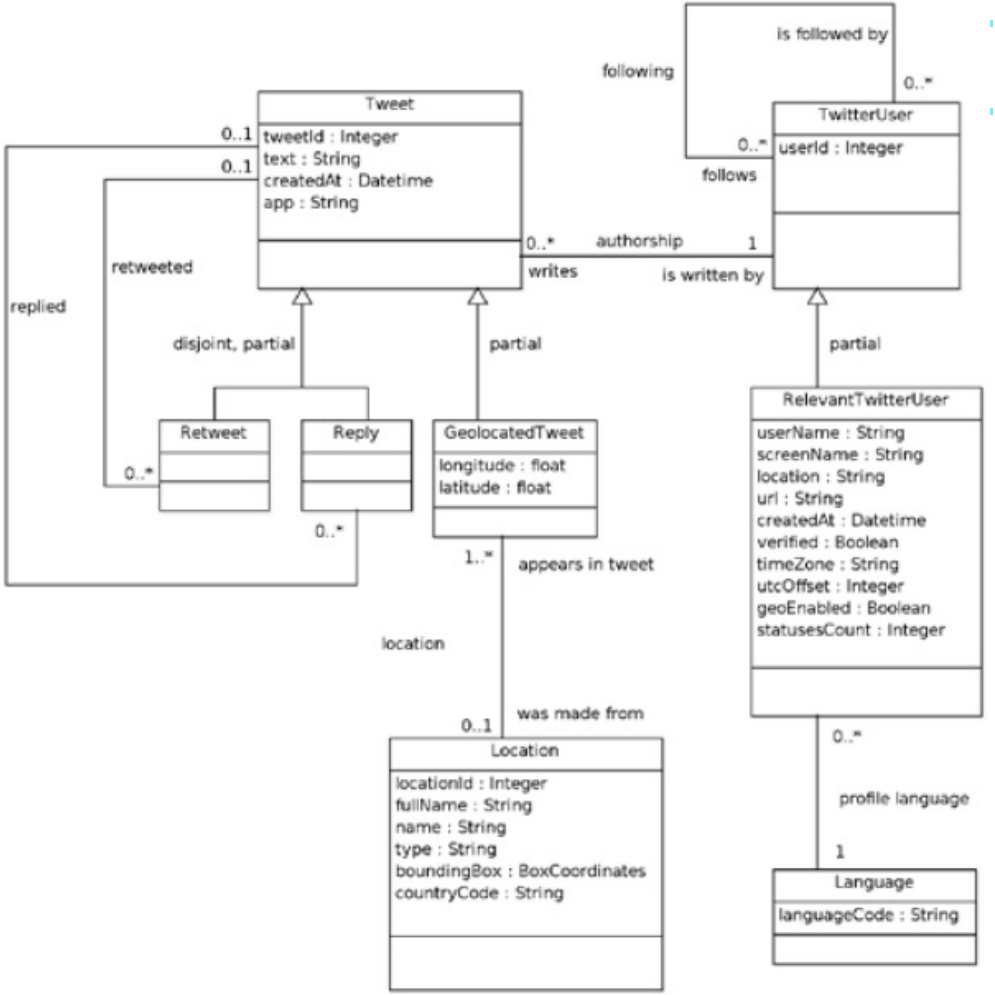
00.1 Captura de datos



Universitat Oberta de Catalunya

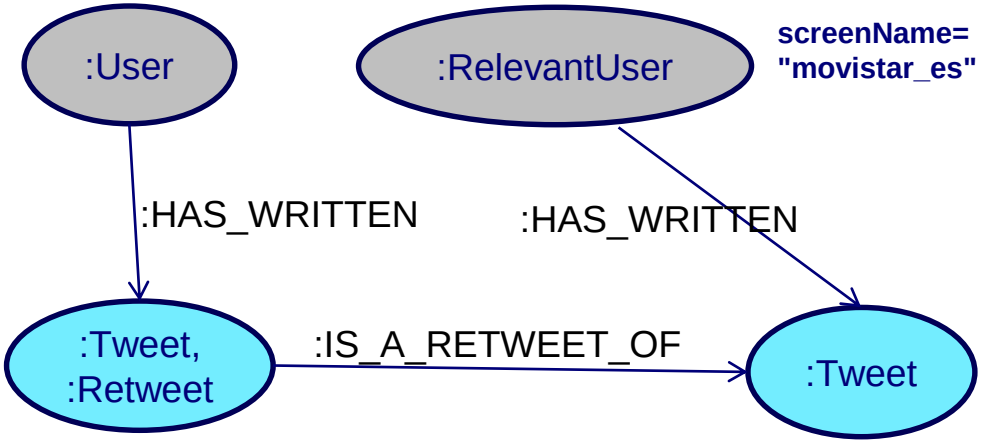
00.1 Esquema conceptual

- Es importante identificar las etiquetas que se utilizarán para identificar los tipos de nodos y aristas.
- Es conveniente que haya una etiqueta para cada tipo de datos sobre el que queramos realizar consultas.
- Se deberán identificar también los atributos para cada tipo y los índices a definir.



00.1 Extracción y carga de datos

Tipo	Nodos
Tweet	> 2.600.000
Reply	> 600.000
Retweet	> 1.600.000
GeoLocatedTweet	> 18.000
User	> 1.200.000
RelevantTwitterUser	499
Language	2
Location	> 2.500



Tipo	Participantes	Número de relaciones
FOLLOWS	TwitterUser--> TwitterUser	705.345
HAS WRITTEN	TwitterUser --> Tweet	2.644.051
HAS_A_PROFILE_LANGUAGE	RelevantTwitterUser --> Language	499
IS_WRITTEN_FROM	GeoLocatedTweet --> Location	16.339
IS_A_REPLY_OF	Reply --> Tweet	577.743
IS_A_RETWEET_OF	Retweet --> Tweet	1.564.319

00.1 Índice

- ¿Qué problema queremos resolver?
- ¿Qué herramientas utilizaremos para hacerlo?
  - ¿Qué es una base de datos NoSQL?
  - ¿Qué es una base de datos en Grafo?
  - ¿Qué es un grafo?
  - ¿Qué es Neo4j?
- ¿Cómo se ha realizado la captura y almacenaje de datos?
- **¿Cómo se ha realizado el análisis de datos?**

00.1 Análisis de datos: Actividad en la red

Los 10 últimos tweets del usuario de twitter de Movistar (screenName= moviestar_es)

```
MATCH (u:RelevantTwitterUser)-[:HAS_WRITEN]->(t:Tweet)
WHERE u.screenName="movistar_es"
RETURN u, t
ORDER BY t.createdAt DESC LIMIT 10
```

El número medio de Tweets por usuario relevante

```
MATCH (u:RelevantTwitterUser)-[:HAS_WRITEN]->(t:Tweet)
WITH u as user, count(t) as num_tweets
RETURN avg(num_tweets) AS average_tweets_x_user
```

00.1 Análisis de datos: Aplicaciones que usan los usuarios

Qué aplicaciones hay y cuantos usuarios han usado cada una de ellas (sólo las 5 más prolíficas)

```
MATCH (t:Tweet)
WITH t.app AS app, count(*) AS numTweetsPerApp
WHERE app<>"web"
RETURN app, numTweetsPerApp
ORDER BY numTweetsPerApp DESC limit 5
```

00.1 Análisis de datos: Detección de usuarios importantes

¿Cuales son los 5 usuarios relevantes con más seguidores?

```
MATCH (:TwitterUser)-[:FOLLOWS]->(u:RelevantTwitterUser)
WITH u AS user, count(*) AS numFollowers
RETURN user.screenName, numFollowers
ORDER BY numFollowers DESC LIMIT 5
```

¿Cuales son los 5 usuarios relevantes con más seguidores de segundo nivel?

```
MATCH (:TwitterUser)-[:FOLLOWS]->(:TwitterUser)-[:FOLLOWS]->
    (u:RelevantTwitterUser)
WITH u AS user, count(*) AS numFollowers
RETURN user.screenName, numFollowers
ORDER BY numFollowers DESC LIMIT 5
```

00.1 Análisis de datos: Detección de usuarios importantes

¿Cuales son los 5 usuarios relevantes con más seguidores de tercer nivel?

```
MATCH (:TwitterUser)-[:FOLLOWS*3]->(u:RelevantTwitterUser)
WITH u AS user, count(*) AS numFollowers
RETURN user.screenName, numFollowers
ORDER BY numFollowers DESC LIMIT 5
```

¿Cuál es el camino de *followers* más corto entre el usuario todo par de usuarios relevantes?

```
MATCH p = shortestPath( (u1:RelevantTwitterUser{screenName:"ictlogist"}
    )-[r:FOLLOWS*]->(u2:RelevantTwitterUser) )
WHERE u1 <> u2
WITH EXTRACT( n IN NODES(p) |n.screenName) AS nodes
RETURN nodes
```

00.1 Análisis de datos: Segmentación de comportamiento

Numero de usuarios en común por cada par de usuarios relevantes

```
MATCH (u1:RelevantTwitterUser)-[:FOLLOWS]->(:TwitterUser)<-
    [:FOLLOWS]-(u2:RelevantTwitterUser)
RETURN u1.screenName, u2.screenName, count(*) AS
    numCommonFriends
ORDER BY count(*) DESC LIMIT 20
```

Conclusiones

- Hay un gran abanico de bases de datos disponibles
- Una base de datos NoSQL en grafo puede ser una buena opción para
    - Representar datos altamente relacionados (redes sociales),
    - Realizar análisis básicos sobre los datos,
    - Interaccionar con los datos para buscar patrones,
    - Cuando queramos aplicar un mismo conjunto de análisis sobre distintos conjuntos de datos
- En caso de necesitar técnicas de análisis más avanzadas, puede ser necesario utilizar otras herramientas. Aún así, la base de datos en grafo podría alojar el almacén de datos analíticos donde se ubican los datos a analizar.

Universitat Oberta
de Catalunya

Espero que haya sido un ejemplo claro y motivador para ver la utilidad y potencia de las bases de datos NoSQL

@Jordi_Conesa
@ jconesac@uoc.edu

UOC.universitat
@UOCuniversitat
UOCuniversitat