Normas

1. La realización de la **práctica es obligatoria** y será la nota final de la asignatura.
2. La práctica está pensada para realizarse en **grupos de 2 a 3 personas**. Excepcionalmente puede realizarse y entregarse de forma individual, previa consulta a los profesores.
3. Antes de la entrega de la práctica, se deberá **notificar** a los dos profesores de la asignatura, a través de la herramienta de correo del Aula Virtual **la composición del grupo**, poniendo en copia del mensaje también a todos los miembros del grupo.
4. La **fecha límite de entrega** es el miércoles, **18 de abril de 2018**. La práctica se entregará a través del Aula Virtual de la asignatura (pestaña Evaluación). Bastará que lo entregue uno de los miembros del grupo. Las notas se publicarán como muy tarde el día 7 de mayo de 2018.
5. La entrega consta de dos partes:
   5.1. Un fichero con la memoria explicativa en formato pdf con el siguiente formato: "**MDS_Memoria_*ApellidoAlumno1YApellidoAlumno2*.pdf"**
   5.2. Un fichero comprimido con el siguiente formato: "**MDS_Codigo_*ApellidoAlumno1YApellidoAlumno2*.zip**" que contendrá todos los ficheros fuente comentados: programa de transformación de XML a formato JSON/CSV/RDF, código de las consultas desarrolladas, etc. También debe incluir un fichero README adicional, indicando los pasos a seguir para su ejecución.

   En ningún caso debe enviar la fuente de datos ni completa ni parcial. En este último caso, simplemente indicará exactamente el subconjunto del que ha partido y cómo lo ha generado (dando los detalles para poder reproducirlo).

Especificación

En esta práctica se partirá de una fuente de datos real, y que se caracteriza por un elevado grado de interdependencia interna. Además, el tamaño de esta fuente (en la forma de un único fichero) hará difícil su manejo, mostrando (aun a pequeña escala) algunas de las dificultades habituales en el procesamiento de esta información.

Contexto: fuente de datos

La fuente de datos es la base de datos DBLP Computer Science Bibliography http://www.informatik.unitrier.de/~ley/db/, considerada como la mayor recopilación existente de referencias bibliográficas académicas específicamente centrada en la informática (en cualquiera de sus variantes). En particular, almacena los datos relativos a la gran mayoría de las revistas científicas y congresos académicos sobre informática, en muchos casos remontándose hasta publicaciones de los años 60.

La recopilación completa de datos se puede descargar como un único fichero XML desde http://dblp.uni-trier.de/xml/. Actualmente (26/02/2018), este fichero comprimido (gzipped) tiene un tamaño de 409 MB y expandido ocupa más de 2 GB. En el documento DBLP - Some Lessons Learned, se describen los detalles de formato -- en particular, la información relativa a los distintos campos para cada tipo de elemento. En la misma dirección se puede encontrar el DTD que define estos elementos. Hay ocho tipos de elementos, de los que se han de destacar tres: artículos de revista (article), artículos en congresos (inproceedings) y artículos en libros (incollection). Para cada artículo se indican, al menos, los autores, el título, las páginas y la fecha, junto a otra información complementaria. El objetivo de este trabajo es utilizar esta información para obtener una serie de medidas que permitirán evaluar a los investigadores en informática.

En el caso de tener dificultades con el tamaño de la fuente de datos, se puede utilizar como simplificación el resultado de una búsqueda en http://www.dblp.org/search, que proporciona un subconjunto de los resultados de la fuente de datos original. En el caso de optar por esta simplificación deberá seleccionar las publicaciones del 2017 (es decir, el criterio de búsqueda será year:2017), que actualmente son unas 277.000. El resultado de la consulta se podrá descargar en diferentes formatos: XML, JSON y JSONP, aunque la API solo permite la exportación de los primeros 1000 resultados. Se **debe optar** por el formato de salida XML. Además debe tenerse en cuenta que la salida en este caso tendrá una estructura diferente que en el caso del dataset completo.

Importante: Se valorará positivamente el procesamiento del dataset completo.

Desarrollo de la Práctica

La práctica tiene 2 partes bien diferenciadas:

Parte I: MongoDB (80%)

1) **Captura y procesamiento de datos.**
   La captura y procesamiento de datos se refiere a descargar el fichero dblp.xml y procesarlo para convertirlo al formato JSON, pudiéndose descartar aquellos elementos que se consideren irrelevantes.
   Para ello, en primer lugar, debe analizar la fuente de datos y el fichero DTD que define estos elementos, así como las consultas más frecuentes a realizar para hacer un diseño del esquema de la Base de Datos de MongoDB.
   Para procesar el fichero XML y realizar la conversión a formato JSON, debe construir un programa en Python. Una estrategia muy práctica, aunque relativamente lenta, consiste en procesar el fichero por "fragmentos" de menor tamaño.
2) **Almacenamiento de datos.**
   Una vez convertidos al formato necesario (y adecuadamente pre-procesados como se prefiera), los datos deben almacenarse en una base de datos MongoDB. En la memoria debe explicar claramente y justificar el diseño de las colecciones implementadas en la base de datos.
3) **Análisis de datos.**

   En esta etapa se deben analizar (consultar) los datos almacenados, usando el lenguaje de consulta propio de la Base de Datos. Se pueden utilizar los medios que se deseen (incluso almacenamiento de datos calculados en la propia BD) con el fin de facilitar la obtención de estas respuestas de la manera más simple posible. En las consultas en las que se haga referencia a un autor determinado, elija uno de los que esté en su conjunto de datos.

   En el caso de usar el dataset simplificado, el resultado obtenido puede no ser realista, dando que solo se trabaja con un año.

   Las consultas son las siguientes:

   1. Listado de todas las publicaciones de un autor determinado.
   2. Número de publicaciones de un autor determinado.
   3. Número de artículos en revista para el año 2017.
   4. Número de autores ocasionales, es decir, que tengan menos de 5 publicaciones en total.
   5. Número de artículos de revista (*article*) y número de artículos en congresos (*inproceedings*) de los diez autores con más publicaciones totales.
   6. Número medio de autores de todas las publicaciones que tenga en su conjunto de datos.
   7. Listado de coautores de un autor (Se denomina coautor a cualquier persona que haya firmado una publicación).
   8. Edad de los 5 autores con un periodo de publicaciones más largo (Se considera la Edad de un autor al número de años transcurridos desde la fecha de su primera publicación hasta la última registrada).
   9. Número de autores novatos, es decir, que tengan una Edad menor de 5 años. Se considera la Edad de un autor al número de años transcurridos desde la fecha de su primera publicación hasta la última registrada.
   10. Porcentaje de publicaciones en revistas con respecto al total de publicaciones.

Defina los índices que crea necesarios para optimizar las consultas, justificando su elección.

Parte II: Neo4J (20%)

Dada la naturaleza de los datos, también resulta especialmente sencillo utilizar una base de datos orientada a grafos, tal como Neo4j.

1) **Captura y procesamiento de datos.**

   Debe procesar el fichero XML del apartado anterior para convertirlo a un formato adecuado.

2) **Almacenamiento de datos.**

   Una vez convertidos al formato necesario (y adecuadamente pre-procesados como se prefiera), los datos deben almacenarse en una base de datos Neo4J.

   En la memoria debe justificar el diseño de la base de datos orientada a grafos que haya elegido.

   Comente brevemente las diferencias que encuentren entre ambos sistemas (MongoDB y Neo4J)

4) **Análisis de datos.**

   Proponga e implemente al menos 3 consultas para las que claramente Neo4j sea más apropiado que MongoDB.

   Defina los índices que considere necesario para optimizar las consultas definidas.