Bases de Datos a Gran Escala
Trabajos

Diego Sevilla Ruiz

Dpto. Ingeniería y Tecnología de Computadores
Facultad de Informática
Universidad de Murcia

dsevilla@um.es

2017

Instrucciones para la realización de los trabajos

- En las siguientes transparencias se incluyen los trabajos propuestos para la asignatura
- Cada trabajo incluye el estudio de una base de datos vista en clase y la importación de los datos de Stackoverflow
- El trabajo será INDIVIDUAL
- Más de un alumno podrá elegir el mismo trabajo, pero como máximo un trabajo podrá ser elegido por dos alumnos
- Se entregará una memoria (y código, si se ha generado) en una tarea abierta a tal fin
- La fecha de entrega será el **día del examen de la asignatura**
- Para apuntarse a un trabajo el alumno tendrá que:

Instrucciones para la realización de los trabajos (cont.)

1. Conectarse al servidor (usuario: alumno, pass: bdge2017)

```
$ mysql -hneuromancer.inf.um.es -P3307 \
    -ualumno -pbdge2017 \
    --default-char-set=utf8mb4 \
    --protocol=tcp trabajos
```

   O bien si se usan los contenedores instalados:

```
$ docker run --rm -ti mysql mysql \
    -hneuromancer.inf.um.es -ualumno \
    -pbdge2017 --protocol=tcp -P3307 \
    --default-character-set=utf8mb4 trabajos
```

2. Añadir una entrada a la tabla asignacion_trabajos con su dni, nombre y id del trabajo elegido

Instrucciones para la realización de los trabajos (cont.)

3. Se pueden listar qué trabajos hay y cuáles quedan libres: (no se muestra el campo spec, que guarda la especificación del trabajo)

```
mysql> select id,titulo from trabajos;
+-----+--------------------------+
| id  | titulo                   |
+-----+--------------------------+
| T01 | Open TSDB                |
| T02 | Apache Cassandra         |
| T03 | OrientDB                 |
| T04 | Redis                    |
| T05 | Elasticsearch            |
| T06 | CouchBase & N1QL         |
| T07 | Riak                     |
| T08 | RethinkDB                |
| T09 | InfluxDB                 |
| T10 | Accumulo                 |
| T11 | ArangoDB                 |
| T12 | Tecnologías Serverless   |
| T13 | Apache Sqoop             |
| T14 | Apache Pig               |
| T15 | CosmosDB                 |
+-----+--------------------------+
```

Instrucciones para la realización de los trabajos (cont.)

```
mysql> select * from asignados;
+-----+---------------------------+-------------+
| id  | titulo                    | nasignados  |
+-----+---------------------------+-------------+
| T01 | Open TSDB                 |           0 |
| T02 | Apache Cassandra          |           0 |
| T03 | OrientDB                  |           0 |
| T04 | Redis                     |           0 |
| T05 | Elasticsearch             |           0 |
| T06 | CouchBase & N1QL          |           0 |
| T07 | Riak                      |           0 |
| T08 | RethinkDB                 |           0 |
| T09 | InfluxDB                  |           0 |
| T10 | Accumulo                  |           0 |
| T11 | ArangoDB                  |           0 |
| T12 | Tecnologías Serverless    |           0 |
| T13 | Apache Sqoop              |           0 |
| T14 | Apache Pig                |           0 |
| T15 | CosmosDB                  |           0 |
+-----+---------------------------+-------------+
```

Instrucciones para la realización de los trabajos (cont.)

- Las tablas disponibles:

```
mysql> show tables;
+------------------------+
| Tables_in_trabajos     |
+------------------------+
| asignacion_trabajos    |
| asignados              |
| trabajos               |
+------------------------+
```

- El código de creación de la base de datos es un ejemplo de creación de usuarios y de permisos, y puede verse en el git de la asignatura aquí: `https://github.com/dsevilla/bdge/blob/17-18/addendum/trabajos/creatrabajos.sql`

T01. Open TSDB

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de series de datos, como si, por ejemplo, comentarios, posts, etc. se ejecutaran en un stream)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura
- `http://opentsdb.net/`

T02. Apache Cassandra

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, lenguaje de consultas CQL, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación siguiendo el modelo de documentos)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T03. OrientDB

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, lenguaje de consultas, grafos vs. documentos, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación y grafos)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T04. Redis

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, *framework* de procesamiento map/reduce, replicación multiservidor, lenguaje de consultas, uso de varias estructuras de datos (listas, mapas), etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de diferentes estructuras de datos)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T05. Elasticsearch

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, organización en etiquetas, búsquedas complejas, replicación multiservidor, lenguaje de consultas, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (organización de etiquetas)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T06. CouchBase & N1QL

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, organización en etiquetas, búsquedas complejas, replicación multiservidor, lenguaje de consultas N1QL, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (documentos y consultas)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T07. Riak

- Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación siguiendo el modelo de documentos)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T08. RethinkDB

- `https://rethinkdb.com/`. Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación donde sea posible)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T09. InfluxDB

- `https://www.influxdata.com/time-series-platform/influxdb/`. Pasos de instalación de la base de datos (a ser posible en la máquina virtual)
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, tratamiento de series temporales, uso del API HTTP, replicación multiservidor, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación donde sea posible)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T10. Accumulo

- `http://accumulo.apache.org/`. Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, tratamiento de columnas, replicación multiservidor, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de columnas)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T11. ArangoDB

- `https://www.arangodb.com/`. Pasos de instalación de la base de datos
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, lenguaje de consultas AQL, grafos vs. documentos, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima (uso de agregación y grafos)
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T12. Tecnologías Serverless

- Pasos de uso de cada plataforma. Al menos: AWS Lambda y Azure Functions (también se puede considerar Google UDF)
- Descripción del modo de funcionamiento, posibilidades de modelado de datos y características
- Mostrar cómo trabajar con los datos de Stackoverflow
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T13. Apache Sqoop

- `http://sqoop.apache.org/`. Pasos de instalación de la herramienta
- Descripción de la herramienta, posibilidades de transformación y carga de datos, modos de funcionamiento, posibilidades de cambio de formato de datos, etc.)
- Mostrar cómo importar los datos de Stackoverflow (de CSV a MySQL, de CSV a HBase, viendo cómo organizar la base de datos)
- API de creación de trabajos *batch*
- Generar código de importación con `sqoop-codegen`

T14. Apache Pig

- Pasos de instalación de la herramienta
- Descripción de la herramienta, posibilidades de transformación y carga de datos, modos de funcionamiento, posibilidades de proceso de datos, etc.
- Mostrar cómo trabajar con los datos CSV de Stackoverflow y mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura

T15. CosmosDB

- Pasos de instalación de la base de datos (o uso en la nube en su caso)
- Descripción de la base de datos, modo de funcionamiento, posibilidades de modelado de datos y características (si permite transacciones, framework de procesamiento map/reduce, replicación multiservidor, lenguaje de consultas, etc.)
- Mostrar cómo importar los datos de Stackoverflow
- Mostrar cómo redistribuir los datos de Stackoverflow de forma óptima
- Mostrar cómo se realizarían las consultas RQ1 a RQ4 de los artículos vistos en la sesión 2
- Realizar pruebas de eficiencia comparada con alguna de las bases de datos vistas en la asignatura