Sequoia User's Guide

Emmanuel Cecchet

Julie Marguerite

Mathieu Peltier

Nicolas Modrzyk

Dylan Hansen

Nuno Carvalho

Version 2.10

Table of Contents

1. Getting Started

1.1. What is Sequoia?

Sequoia is a database cluster middleware that allows any Java™ application (standalone application, servlet or EJB™ container, ...) to transparently access a cluster of databases through JDBC™. You do not have to modify client applications, application servers or database server software. You just have to ensure that all database accesses are performed through Sequoia.

Sequoia is a *free*, *open source* project that is the continuation of the C-JDBC project (http://c-jdbc.objectweb.org) hosted by the ObjectWeb Consortium (http://www.objectweb.org/). Sequoia is licensed under an Apache v2 license (http://www.apache.org/licenses/LICENSE-2.0.html) is licensed whereas C-JDBC is available under the GNU Lesser General Public License (http://www.gnu.org/copyleft/lesser.html) (LGPL).

Sequoia also provides driver for non-Java applications. These developments are hosted in the Carob project (http://carob.continuent.org). An Eclipse plug-in for Sequoia is also available in the Oak project (http://oak.continuent.org).

1.2. What do I need to use Sequoia?

In order to use Sequoia, you will need:

- a client application that accesses a database through JDBC,
- a JDK™ 1.4 (or greater) compliant Java Virtual Machine™ (JVM)[1],
- a database with a JDBC driver (type 1, 2, 3 or 4) or an ODBC driver used with the JDBC-ODBC bridge.
- a network supporting TCP/IP communications between your cluster nodes.

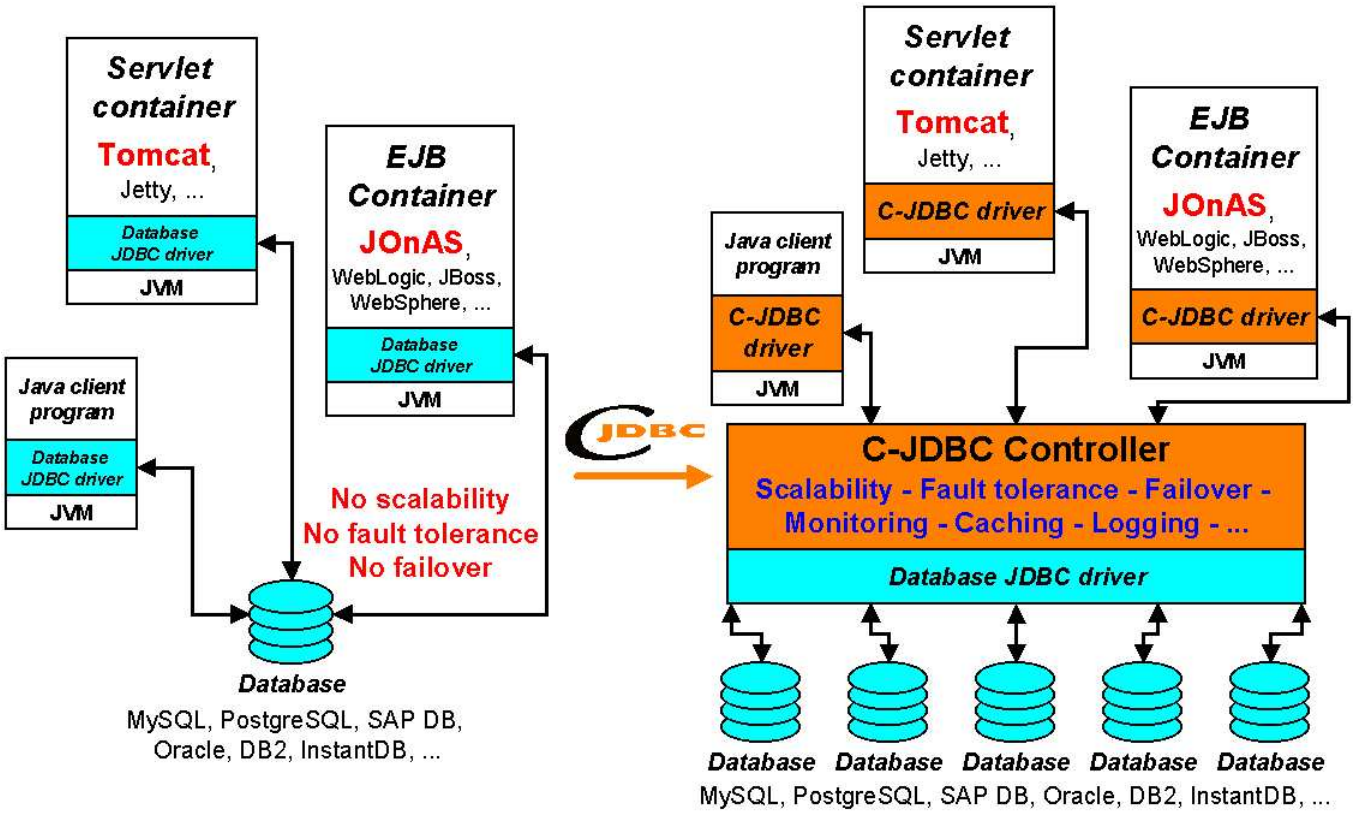
> **Note:** If your client application does not use JDBC, you can use the C++ API or the ODBC driver provided by the Carob project (http://carob.continuent.org/).

1.3. Why should I use Sequoia?

You have a Java application or a Java-based application server that accesses one or several databases. The database tier becomes the bottleneck of your application or it is a single point of failure or both. Sequoia can help you resolve these problems by providing:

- performance scalability by adding database nodes and balancing the load among these nodes.
- high availability of the database tier, i.e. Sequoia tolerates database crashes and offers transparent failover using database replication techniques.
- improved performance with fine grain query caching and transparent connection pooling.
- SQL traffic logging for performance monitoring and analysis.
- support for clusters of heterogeneous database engines.

Figure 1. Sequoia principle



1.4. How does it work?

Sequoia provides a flexible architecture that allows you to achieve scalability, high availability and failover with your database tier. Sequoia implements the concept of RAIDb: *Redundant Array of Inexpensive Databases* (see Section 9). The database is distributed and replicated among several nodes and Sequoia load balances the queries between these nodes.

Sequoia provides a generic JDBC driver to be used by the clients (see Section 5). This driver forwards the SQL requests to the Sequoia controller (see Section 7) that balances them on a cluster of databases (reads are load balanced and writes are broadcasted). Sequoia can be used with any RDBMS (Relational DataBase Management System) providing a JDBC driver, that is to say almost all existing open source and commercial databases. Figure 1 gives an overview of the Sequoia principle.

Sequoia allows to build any cluster configuration including mixing database engines from different vendors. The main features provided by Sequoia are performance scalability, fault tolerance and high availability. Additional features such as monitoring, logging, SQL requests caching are provided as well.

The architecture is widely open to allow anyone to plug custom requests schedulers, load balancers, connection managers, caching policies, ...

1.5. What does it cost?

From a software point of view, Sequoia is an open-source software licensed under Apache v2 License which means that it is free of charge for any usage (personal or commercial). If you are using commercial RDBMS

(such as Oracle, DB2, ...), you will have to buy extra licenses for the nodes where you install replicas of the database. But you can possibly use open-source databases to host replicas of your main database.

You need to buy extra machines if you want more performance and more fault tolerance. Sequoia has been designed to work with standard off-the-shelf workstations because it primarily targets low cost open-source solutions but it can work as well with large SMP machines. A standard Ethernet network is sufficient to achieve good performance.

1.6. What kind of modifications are needed?

You *do not have to change anything* to your application or your database.

You only have to update the JDBC driver configuration used by your application (usually it is just a configuration file update) and to setup a Sequoia configuration file (see Section 10).

2. Getting the Software

The binary distribution of Sequoia can be downloaded from Sequoia's Web site (http://sequoia.continuent.org/). It mainly contains the JAR files for the Sequoia driver and controller and also the documentation and other tools such as the Sequoia administration console.

> **Note:** A source distribution of Sequoia is also available. The whole code base can also be downloaded through an anonymous CVS server[2]. For more information, please refer to Sequoia Developer's Guide. Most users will only need the binary distribution.

The following formats are available (where `x.y` is the Sequoia release number):

- `sequoia-x.y-bin-installer.jar`: Java graphical installer (powered by IzPack (http://www.izforge.com/izpack/)).
- `sequoia-x.y-bin.tar.gz`: binary distribution for the Unix platforms users.
- `sequoia-x.y-bin.zip`: binary distribution for the Windows platforms users.

We strongly advice to use the Java installer package since it automatically configures the scripts to suit your system configuration.

> **Note:** All distributions contain the user documentation.

3. Installation

3.1. Sequoia Controller

3.1.1. Using the Java graphical installer

The easiest way to install Sequoia is to use the Java graphical installer. A Java Virtual Machine is of course needed in this case.

- Unix users can simply launch the installation program by typing:

```
bash> java -jar sequoia-x.y.bin-installer.jar
```

- Windows users can use the same command or just double-click on the JAR installation file if your JRE has been properly installed.

3.1.2. Using the binary distribution

If you want to use the other distribution formats (for example if you have not installed a JVM or if you can not launch a graphical application), you have to uncompress the downloaded file in the directory of your choice, and then set the `SEQUOIA_HOME` environment variable.

> **Note:** If you are using the Java installer, you do not need to set any environment variable since the installer customizes the scripts with the installation path.

To set the `SEQUOIA_HOME` environment variable, you can proceed as follows:

- Unix users can proceed as follows:

```
bash> mkdir -p /usr/local/sequoia
bash> cd /usr/local/sequoia
bash> tar xfz /path-to-sequoia-bin-dist/sequoia-x.y-bin.tar.gz
bash> export SEQUOIA_HOME=/usr/local/sequoia
```

> **Note:** In this example, we assume you install Sequoia in the `/usr/local/sequoia` directory.

  You can modify your shell configuration file (`.bashrc`, `.cshrc`, ...) to set the environment variable permanently.

- Windows users have to use an utility such as WinZip (http://www.winzip.com/) to extract the files from the archive. Then, to set the `CJDB_HOME` variable, do the following according to your Windows version:
  - *Windows 95 or 98*: you must insert the following line in the `AUTOEXEC.BAT` file:

```
set SEQUOIA_HOME="C:\Program Files\Sequoia"
```

- *Windows Me*: go to the "Start Menu", then choose "Programs", "Accessories", "System Tools" and "System Information". A window titled "Microsoft Help and Support" should appear. Select the "Tools" menu, and choose the "System Configuration Utility". Go to the "Environment" and click on the "New" button. Enter `SEQUOIA_HOME` in the "Variable Name" field and `"C:\Program Files\Sequoia"` in "Variable Value". Once you have changed and saved the value, you will be prompted for reboot.
- *Windows NT*: go to the "Start Menu", then choose "Settings", "Control Panel" and select "System". Select the "Environment" tab and click on the "New" button. Enter `SEQUOIA_HOME` in the "Variable Name" field and `"C:\Program Files\Sequoia"` in "Variable Value".
- *Windows 2000*: go to the "Start Menu", then choose "Settings", "Control Panel" and select "System". Select the "Advanced" tab and click on the "New" button. Enter `SEQUOIA_HOME` in the "Variable Name" field and `"C:\Program Files\Sequoia"` in "Variable Value".
- *Windows XP*: go to the "Start Menu", then double click on "System". In the "System Control Panel" select the "Advanced" tab and push the `Environment Variables` button. Click on the "New" button for "System Variables". Enter `SEQUOIA_HOME` in the "Variable Name" field and `"C:\Program Files\Sequoia"` in "Variable Value".

Note: In this example, we assume you install Sequoia in the `C:\Program Files\Sequoia` directory.

Note: Do not forget the quotes in the SEQUOIA_HOME environment variable definition else the starting scripts will fail with paths including spaces.

3.2. Sequoia Driver

Once you have installed the Sequoia controller, you will find the driver JAR file in the `drivers/` directory of the controller installation location.

To install the Sequoia driver, you just have to add the `sequoia-driver.jar` file to the client application classpath. This driver replaces the database native driver in the client application. The database native driver will be used by the Sequoia controller to access your database. Therefore, the Sequoia driver and controller can be seen as a proxy between your application and your database native driver.

3.3. Sequoia out of the box

A demo featuring a RAIDb-1 configuration of HyperSonic SQL databases can be started by launching the **demo-raidb1.sh** or **demo-raidb1.bat** file from the `demo` directory in your Sequoia installation.

This is especially useful if you are new to clustering, or new to Sequoia. The setup used is as follows:

- 2 HyperSonic SQL databases are started on two different ports (9001 and 9002)
- An extra HyperSonic SQL database is started on port 9003 to be used as the recovery log database
- The Sequoia controller is configured to load automatically a virtual database containing those two HyperSonic SQL backends. The controller startup configuration file is found in

`SEQUOIA_HOME/config/controller/controller-raidb1.xml` and the virtual database configuration file is `SEQUOIA_HOME/config/virtualdatabase/hsqldb-raidb1.xml`.

- Once the RAIDb-1 configuration is loaded, you can connect to Sequoia using iSQL, a graphical SQL console bundled with Sequoia. You can start iSQL by using **isql.sh** or **isql.bat**.

The login to use for Sequoia is `user` with an empty password. The login for both HSQL databases is `test` with an empty password.

Note: A tutorial in the documentation section of the Sequoia web site describes the usage of this demo.

4. Migrating from C-JDBC to Sequoia

The C-JDBC name had to be changed due to Sun?s trademark of JDBC. Therefore, C-JDBC now becomes Sequoia. Sequoia is the continuation of C-JDBC and builds upon the same code base, so the migration should be straightforward for current users.

4.1. What is new with Sequoia?

Sequoia is now backed by a team of 8 full-time engineers working on improving the technology and supporting the community. Our mission is to build industrial quality open source technology. Among the new major open source additions you will find a C++ API and an ODBC driver for non-Java clients as well as a powerful Eclipse plug-in for the management console. The core C-JDBC technology is also greatly improved with a complete redesign of transaction scheduling for a better write parallelism, along with an upcoming, completely rewritten documentation.

4.1.1. Licensing

C-JDBC is distributed under an LGPL open source license. While we like the spirit of the license and we think that an open source community can only grow if we contribute modifications back to the community, the LGPL is hard to enforce in practice and it is not always well understood by users who confuse it with the GPL. Therefore, the contributors and INRIA ,who is the main copyright holder, have agreed to re-license the code under an Apache version 2 license. This will ease code re-use and facilitate contributions for everybody in the community.

4.1.2. Continuent.org (http://continuent.org)

You might wonder why we have setup a new portal? The reason is mainly to offer a better support infrastructure to the community. Continuent.org projects uses a newer version of GForge and integrates JIRA for issue tracking. With JIRA, a more comfortable and flexible system, you can watch the progress of issues, vote on their resolution, see the project roadmap and so on. On Continuent.org, it is also much more flexible to start or host new projects related to the C-JDBC/Sequoia technology. We can easily host more external contributions or projects related to the technology. Don't hesitate to send your contributions! The current projects on Continuent.org are (we are very much in the tree names for Continuent projects!):

- `Sequoia (http://sequoia.continuent.org)`: the continuation of the C-JDBC project including the JDBC driver, core controller, text management console, documentation, ...

- `Hedera (http://hedera.continuent.org)`: a replacement for Tribe that provides a more modular wrapping of group communications so that you have more choices than just JGroups. Other group communication libraries such as Appia are already available with Hedera.
- `Carob (http://carob.continuent.org)` : a C++ client library and API that implements the C-JDBC/Sequoia protocol with the controller and an ODBC driver for Sequoia
- `Oak (http://oak.continuent.org)` : the Eclipse plug-in that replaces the obsolete C-JDBC Swing management console
- `Appia (http://appia.continuent.org)`: Appia is a layered communication framework implemented by the University of Lisbon and providing extended configuration and programming possibilities. Appia is composed by (1) a core that is used to compose protocols and (2) a set of protocols that provide group communication, ordering guarantees, atomic broadcast, among other properties.

4.1.2.1. C-JDBC and Sequoia

So what will happen with C-JDBC? We are committed to support the technology and we will continue to support the community either through c-jdbc@objectweb.org or sequoia@continuent.org. The C-JDBC LGPL code will remain on ObjectWeb and the Sequoia APLv2 code will be on Continuent. Bugs reported on either side will be backported in best effort mode as we always did.

4.2. Migrating your C-JDBC configuration to Sequoia

Here are the steps to migrate your configuration from C-JDBC 2.0.2 to Sequoia 2.2:

1. Copy your `controller.xml` file from the C-JDBC `config/controller` directory to Sequoia `config/controller` directory
2. Rename all instances of "C-JDBC" to "SEQUOIA". Be sure to update DOCTYPE
3. Copy all virtual database configuration files from the C-JDBC `config/virtualdatabase` directory to Sequoia `config/virtualdatabase` directory
4. Rename all instances of "C-JDBC" to "SEQUOIA". Be sure to update DOCTYPE
5. Update the path of the backupers from org.objectweb.cjdbc.controller.backup.OctopusBackuper to org.continuent.sequoia.controller.backup.backupers.OctopusBackuper
6. If using Distribution, add <MessageTimeouts/>
7. Copy contents of drivers directory (except c-jdbc-driver.jar)
8. Any manual changes to jgroups.xml should now be made to total-token.xml

5. Sequoia Driver

5.1. Overview

The Sequoia driver is a generic JDBC driver that is designed to replace any database specific JDBC driver that could be used by a client. The client only has to know on which node the Sequoia controller is running and the name of the database to access. The Sequoia driver implements most of the JDBC 3.0 interface.

Users reported successful usage of Sequoia with the following RDBMS: Oracle®, PostgreSQL, MySQL, Apache Derby, IBM DB2®, Sybase®, SAP DB (MySQL MaxDB), HyperSonic SQL, Firebird, MS SQL Server and InstantDB.

5.2. Loading the Driver

The Sequoia driver can be loaded as any standard JDBC driver from the client program using:

```
Class.forName("org.continuent.sequoia.driver.Driver");
```

Note: The `sequoia-driver.jar` file must be in the client classpath else the driver will fail to load.

5.3. Sequoia JDBC URL

The JDBC URL expected for the use with Sequoia is the following: `jdbc:sequoia://host1:port1,host2:port2/database`.

`host` is the machine name (or IP address) where the Sequoia controller is running, `port` is the port the controller is listening for client connections.

At least one host must be specified but a list of comma separated hosts can be specified. If several hosts are given, one is picked up randomly from the list. If the currently selected controller fails, another one is automatically picked up from the list.

The port is optional is the URL and the default port number is 25322 if it is omitted. Those two examples are equivalent:

```
DriverManager.getConnection("jdbc:sequoia://localhost/tpcw");
DriverManager.getConnection("jdbc:sequoia://localhost:25322/tpcw");
```

Examples using two controllers for fault tolerance:

```
DriverManager.getConnection("jdbc:sequoia://c1.continuent.org,c2.objectweb.org/tpcw");
DriverManager.getConnection("jdbc:sequoia://localhost,remote.continuent.org:2048/tpcw");
DriverManager.getConnection("jdbc:sequoia://smpnode.com:25322,smpnode.com:1098/tpcw");
```

5.3.1. URL options

The Sequoia driver accepts additional options to override the default behavior of the driver. The options are appended at the end of the Sequoia URL after a question mark followed by a list of ampersands separated options. Here is an example:

```
DriverManager.getConnection("jdbc:sequoia://host/db?user=me&password=secret")
```

Another option is to use semicolons to delimit the start of options and options themselves. Example:

```
DriverManager.getConnection("jdbc:sequoia://host/db;user=me;password=secret")
```

The recognized options are:

- `connectionPooling`: By default the Sequoia driver does transparent connection pooling on your behalf meaning that when connection.close() is called, the connection is not physically closed but rather put in a pool for reuse within the next 5 seconds. Set this to false if you do not want the driver to perform transparent connection pooling.

- `debugLevel`: Debug level that can be set to 'debug', 'info' or 'off' to display driver related information on the standard output. Default is off.

- `escapeBackslash`: Set this to false if you don't want to escape backslashes when performing escape processing of PreparedStatements, default is true.

- `escapeSingleQuote`: Set this to false if you don't want to escape single quotes (') when performing escape processing of PreparedStatements, default is true

- `escapeCharacter`: Character to prepend and append to the String values when performing escape processing of PreparedStatements, default is a single quote.

- `user`: user login

- `password`: user password

- `preferredController`: defines the strategy to use to choose a preferred controller to connect to.

  `jdbc:sequoia://node1,node2,node3/myDB?preferredController=ordered` : Always connect to node1, and if not available then try to node2 and finally if none are available try node3.

  `jdbc:sequoia://node1,node2,node3/myDB?preferredController=random`: Pickup a controller node randomly (default strategy)

  `jdbc:sequoia://node1,node2:25343,node3/myDB?preferredController=node2:25343,node3` : Round-robin between node2 and node3, fallback to node1 if none of node2 and node3 is available.

  `jdbc:sequoia://node1,node2,node3/myDB?preferredController=roundRobin`: Round robin starting with first node in URL.

- `retryIntervalInMs`: once a controller has died, the driver will try to reconnect to this controller every retryIntervalInMs to see if the backend is back online. The default is 5000 (5 seconds).

5.4. Getting a connection using a data source

Another way to use the Sequoia driver is to use its `DataSource` implementation. Data sources have been introduced in JDBC 2.0 Standard Extension API and are also a part of JDBC 3.0. They use the Java Naming and Directory Interface (JNDI) to break the application dependence on the JDBC driver configuration (i.e., driver class name, machine name, port number, etc.). With a data source, the only thing an application has to know is the name assigned to the `DataSource` object in the `jdbc` naming subcontext of the JNDI namespace.

The example below registers a data source object with a JNDI naming service. It is typically used by an application server.

```
import org.continuent.sequoia.driver.DataSource;
import javax.naming.Context;
```

```
import javax.naming.InitialContext;
import javax.naming.NamingException;
...
private final static String NAME = "jdbc/sequoia";
private final static String URL = "jdbc:sequoia://localhost:25322/mysql";

// Initializing data source
DataSource ds = new DataSource();
ds.setUrl(URL);

// Get initial context
Context ctx;
try {
  ctx = new InitialContext();
} catch (javax.naming.NamingException _e) {
  ... // Naming exception
}

// Bind data source to a JNDI name
try {
  ctx.bind(NAME, ds);
} catch (javax.naming.NamingException _e) {
  ... // Naming exception
}
```

The `org.continuent.sequoia.driver.DataSource` class implements the `javax.sql.DataSource` JDBC 3.0 interface. The `setUrl` line initializes the data source properties (the URL in this case). The data source object is bound to a logical JNDI name by calling `ctx.bind()`. In the example above, the JNDI name specifies a "jdbc" subcontext and a "sequoia" logical name within this subcontext.

Once a data source object is registered to JNDI, it can be used by an application. The example below gets the data source using the JNDI naming service. Such a piece of code is typically a part of an application that uses JDBC.

```
import javax.naming.Context;
import javax.naming.InitialContext;
import javax.naming.NamingException;
import java.sql.Connection;
import javax.sql.DataSource;
...
private final static String NAME = "jdbc/sequoia";

// Lookup for the data source object
try {
  Context ctx = new InitialContext();
  Object obj = ctx.lookup(NAME);
  if (null == obj) {
    ... // Something wrong: NAME not found
  }
  ctx.close( );
} catch (javax.naming.NamingException _e) {
  ... // Naming exception
}

// Get a new JDBC connection
try {
  DataSource ds = (DataSource) obj;
```

```
  Connection conn = ds.getConnection("user", "sequoia");
  ... // Use of the connection retrieved
  ...
} catch (SQLException _e) {
  ... // SQL exception
}
```

The `ctx.lookup()` line in the example uses the retrieved initial JNDI naming context to do a lookup using the data source logical name. The method returns a reference to a Java object which is then narrowed to a `javax.sql.DataSource` object. Such an object can be then used to open a new JDBC connection by invoking one of its `getConnection()` methods. The application code is completely independent of the driver details, such as the `Driver` class name, URL, etc. (the user name and password used by the connection can be also set by the application server - look at the Sequoia javadoc documentation for more details). The only information a JDBC application has to know is the logical name of the data source object to use.

> **Note:** The URL used for the Sequoia data source is the same as for the `Driver` described in the previous section.

5.5. Stored procedures

Stored procedures are supported by Sequoia since version 1.0b6. Note that Sequoia only support calls in the form **{call <procedure-name>[<arg1>,<arg2>, ...]}** but does not support **{? = call <procedure-name>[<arg1>,<arg2>, ...]}**.

A call to a stored procedure is systematically broadcasted to all backends since there is no way to know if the stored procedure will update the database or not. Therefore, the query cache (see Section 10.6.3), is completely flushed on every stored procedure call. To prevent cache flushing, the user can force the connection to read-only before calling the stored procedure. But never set a connection to read-only when calling a stored procedure that updates the database. If Sequoia detects a read-only connection, it will not flush the cache. However, the call will still be broadcasted to all nodes resulting in duplicated jobs on each backend. Here is an example on how to prevent cache flushing when calling a stored procedure that does only read-only:

```
...
CallableStatement cs = connection.prepareCall("{call myproc(?)}");
cs.setString(1,"parameter1");
// Force no cache flush
connection.setReadOny(true);
// Call the stored procedure without flushing the cache ...
ResultSet rs = cs.executeQuery();
```

In the case of horizontal scalability, only read-only stored procedures are not broadcasted. All other stored procedures returning an int or a ResultSet are executed by all backends at all controllers.

> **Note:** It is not allowed to set a connection to read-only in the middle of a transaction. If you need to set a connection to read-only, you must do so before starting the transaction.

5.6. Blobs: Binary Large Objects

You should not have to change your code for storing blobs into your database. Sequoia will transparently encode the blob in the protocol and forward it to your database driver.

- The column type used to store large objects with `MySQL` is `text`.
- The column type used to store large objects with `PostgreSQL` is `bytea`.

Please refer to the following lines of code for storing and retrieving of large objects:

```
// In the code below:
// The signature of the readBinary method is:
// byte[] readBinary(File file) throws IOException
// it just read a file, and convert its content into an array of bytes

// Store file in database
File fis = new File(storeFile);
query = "insert into ... values(...,?)";
ps1 = con.prepareStatement(query);
if (callBlobMethods)
{
  org.continuent.sequoia.common.protocol.ByteArrayBlob bob =
    new org.continuent.sequoia.common.protocol.ByteArrayBlob(readBinary(fis));
  ps1.setBlob(1, bob);
}
else
{
  ps1.setBytes(1, readBinary(fis));
}
ps1.executeUpdate();
// Read File from database
query = "select * from ... where id=...";
ps1 = con.prepareStatement(query);
ResultSet rs = ps1.executeQuery();
rs.first();
byte[] lisette;
if (callBlobMethods)
{
  Blob blisette = rs.getBlob("blobcolumnname");
  lisette = blisette.getBytes((long) 1, (int) blisette.length());
}
else
{
  lisette = rs.getBytes("blobcolumnname");
}
```

5.7. Clobs: Character Large Objects

CLOB is a built-in type that stores a Character Large Object as a column value in a row of a database table. By default drivers implement Clob using an SQL locator (CLOB), which means that a Clob object contains a logical

pointer to the SQL CLOB data rather than the data itself. A Clob object is valid for the duration of the transaction in which it was created.

Clobs in Sequoia are handled like strings. You can refer to the section of code below to make good usage of clobs. This code is part of the Sequoia test suite.

```
String clob = "I am a clob";
ps = con.prepareStatement("insert into ... values(...,?)");
ps.setString(1, clob);
ps.executeUpdate();

// Test retrieval
String ret;
ps = con.prepareStatement("Select * from ... where id=...");
rs = ps.executeQuery();
rs.first();
clob = rs.getClob("name");
ret = clob.getSubString((long) 0, (int) clob.length());
```

5.8. ResultSet streaming

In its default mode, when a query is executed on a backend, Sequoia makes a copy of the backend's native ResultSet into a Sequoia serializable ResultSet. If the result contains many rows or very large objects, the controller might run out of memory when trying to copy the whole ResultSet.

It is possible to fetch ResultSets by blocks using the Statement.setFetchSize(int rows) method. In this case, the ResultSet will be copied by block of rows and returned when needed by the client. Note that the current implementation only allows to fetch forward streamable ResultSet, which basically means that you are only allowed to call ResultSet.next() on a streamable ResultSet.

Sequoia will try to call setFetchSize() on the backend's driver to let the backend driver also perform the necessary optimizations. However, some driver requires a prior call to setCursorName() in which case you will also have to call setCursorName() on Sequoia to pass it to the backend's driver.

A typical usage of the ResultSet streaming feature is as follows:

```
...
Connection con = getSEQUOIAConnection();
con.setAutocommit(false);
Statement s = con.createStatement();
s.setCursorName("cursor name");
s.setFetchSize(10);
rs = s.executeQuery(sql);
while (rs.next())
{ // Every 10 calls, Sequoia will transfer a new block of rows
  XXX o = rs.getXXX("some column name");
}
...
con.commit();
```

Note: Streamable ResultSets are not cacheable. The result cache automatically detects this kind of ResultSet and does not keep them in the cache. However, as database specific ResultSets are copied into Sequoia ResultSets, the memory footprint of the fetched blocks will be twice the one obtained without

Sequoia. If you have memory restrictions, you can reduce your fetch size by half to reduce the memory footprint of streamed ResultSets.

Streamable ResultSets do not work properly in autocommit mode as the connection used for retrieving the ResultSet is handed back to the pool. The workaround is to always encapsulate the query in a transaction. Note that databases such as PostgreSQL do not support streamable ResultSets in autocommit mode as well.

5.9. Current Limitations

The Sequoia driver currently does not support the following features:

- java.sql.Array and java.sql.Ref types,
- Custom type mapping using java.sql.Connection.setTypeMap(java.util.Map map),
- `XAConnections` (look at the XAPool project (http://xapool.experlog.com) for XA support with Sequoia),
- Streamable ResultSets do not work in autocommit mode.

6. Configuring Sequoia with 3rd party software

6.1. Forenotes on configuring Sequoia with your application

If the application you are using Sequoia with requires a mapper, the best thing to do is to configure the mapping to be that of Sequoia's underlying databases. For example, if you were using JBoss with PostgreSQL , then using Sequoia on top of the PostgreSQL backends with JBoss would imply to still use the mapping for PostgreSQL while plugging the application server to Sequoia (using Sequoia's driver and Sequoia's url).

6.2. Configuring Sequoia with Jakarta Tomcat

Copy the `sequoia-driver.jar` file to the `lib` directory of your web application (for example: `$TOMCAT_HOME/webapps/mywebapp/WEB-INF/lib`).

There are many ways to obtain connections from a Tomcat application. Just ensure that you are using `org.continuent.sequoia.driver.Driver` as the driver class name and that the JDBC URL is a Sequoia URL (see Section 5.3).

6.3. Configuring Sequoia with JOnAS

The `sequoia-driver.jar` file must be found in the JOnAS CLASSPATH.

Here is an example of a `sequoia.properties` file to store in JONAS 3.x `conf` directory (use the `config` directory for JOnAS 2.x):

```
###################### Sequoia DataSource configuration example #
datasource.name      jdbc_1
datasource.url       jdbc:sequoia://someMachine/someDatabase
datasource.classname org.continuent.sequoia.driver.Driver
```

```
datasource.username  your-username
datasource.password  your-password
```

6.4. Configuring Sequoia with JBoss

Copy the `sequoia-driver.jar` file to `$JBOSS_DIST/server/default/lib` for JBoss 3.x or to `$JBOSS_DIST/jboss/lib/ext` for JBoss 2.x.

Here is an example of a datasource configuration file to be used with JBoss:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- ===================================================================== -->
<!--                                                                       -->
<!-- JBoss Server Configuration                                            -->
<!--                                                                       -->
<!-- ===================================================================== -->

<!-- ===================================================================== -->
<!-- Datasource config for Sequoia                                          -->
<!-- ===================================================================== -->
<datasources>
  <local-tx-datasource>
    <jndi-name>sequoia-DS</jndi-name>
    <connection-url>jdbc:sequoia://localhost:25322/lscluster</connection-url>
    <driver-class>org.continuent.sequoia.driver.Driver</driver-class>
    <user-name>user</user-name>
    <password>tagada</password>
  </local-tx-datasource>
</datasources>
```

6.5. Configuring Sequoia with BEA Weblogic Server 7.x/8.x

Place the `sequoia-driver.jar` file in the classpath of the Weblogic Server.

Here is an example of a connection pool configuration for use with Weblogic:

```
<JDBCConnectionPool
    DriverName="org.continuent.sequoia.driver.Driver"
    InitialCapacity="1"  MaxCapacity="15"
    Name="sequoiaPool" Properties="user=username;password=password"
    ShrinkingEnabled="true" SupportsLocalTransaction="true"
    Targets="wlservername" URL="jdbc:sequoia://192.168.0.1/vdb"
    XAPreparedStatementCacheSize="0"/>
```

Next, create the required `TXDataSources`:

```
<JDBCTxDataSource EnableTwoPhaseCommit="true"
  JNDIName="sequoia-DS" Name="Sequoia TX Data Source"
  PoolName="sequoiaPool" RowPrefetchEnabled="true" Targets="wlservername"/>
```

6.6. Configuring Sequoia with Hibernate

Sequoia just has to be defined as any JDBC driver in Hibernate, leaving the syntax set to the proper database. Here is a configuration example to use Hibernate with a Sequoia cluster made of Sybase backends:

```
## Sequoia
hibernate.dialect                net.sf.hibernate.dialect.SybaseDialect
hibernate.connection.driver_class org.continuent.sequoia.driver.Driver
hibernate.connection.username     user
hibernate.connection.password     pass
hibernate.connection.url          jdbc:sequoia://localhost:25322/test
```

6.7. Using sequences with Hibernate, Sequoia and PostgreSQL

Our Hibernate dialect is as follows:

```
import net.sf.hibernate.dialect.PostgreSQLDialect;
public class SEQUOIAPostgreSQLDialect extends PostgreSQLDialect
{
  public String getSequenceNextValString(String sequenceName)
  {
    return "{call nextval('"+sequenceName+"')}";
  }
}
```

We simply extend the default PostgreSQL Dialect and override the getSequenceNextValString() method and tell it to use "{call ..." so that all the sequences in the cluster get incremented.

We then changed our Hibernate conf file to user to our custom dialect instead of net.sf.hibernate.dialect.PostgreSQLDialect.

7. Sequoia controller

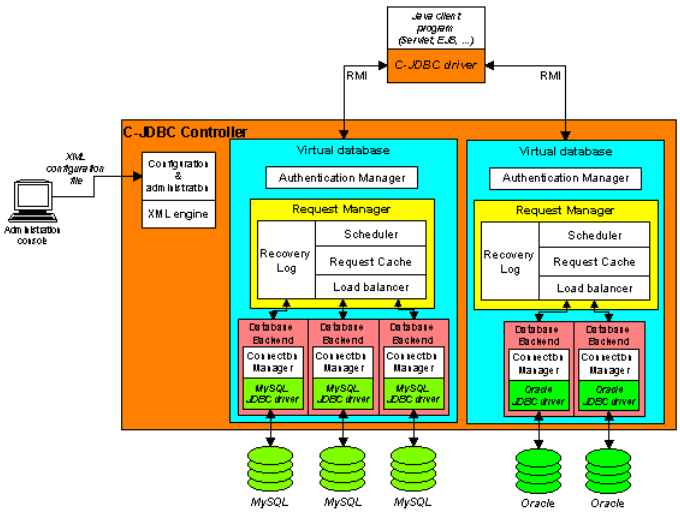
7.1. Design Overview

The Sequoia controller is made of several components as shown in Figure 2. The controller hosts *virtual databases*. A *virtual database* gives the illusion of a single database to the user. It exports the same database name and login/password as those used in the client application. Therefore the client application can run unmodified with Sequoia.

When the client application connects to the database using an URL like `jdbc:sequoia://host:25322/myDB`, the Sequoia driver tries to connect to a Sequoia controller running on port `25322` on node `host`. Once the connection is established the login and password are sent with the `myDB` database name to be checked by the controller.

A virtual database contains the following components:

Figure 2. Sequoia controller design overview



- *authentication manager*: it matches the virtual database login/password (provided by the application to the Sequoia driver) with the real login/password to use on each backend. The authentication manager is only involved at connection establishment time.
- *backup manager*: manages a list of generic or database specific Backupers that are in charge of performing database dump and restore operation. Backupers should also take care of transferring dumps from one controller to another.
- *request manager*: it handles the requests coming from a connection with a Sequoia driver. It is composed of several components:
  - *scheduler*: it is responsible for scheduling the requests. Each RAIDb level has its own scheduler.
  - *request caches*: these are optional components that can cache query parsing, the result set and result metadata of queries.
  - *load balancer*: it balances the load on the underlying backends according to the chosen RAIDb level configuration.
  - *recovery log*: it handles checkpoints and allows backends to dynamically recover from a failure or to be dynamically added to a running cluster.

- *database backend*: it represents the real database backend running the RDBMS engine. A *connection manager* mainly provides connection pooling on top of the database JDBC native driver.

Each virtual database and its components are configured using an XML configuration file that is sent from the administration console to the Sequoia controller.

> **Note:** A research report details RAIDb and C-JDBC implementation (http://c-jdbc.objectweb.org/current/doc/RR-C-JDBC.pdf). Other documents and presentations about C-JDBC can be found in the documentation section of the web site (http://c-jdbc.objectweb.org/doc).

7.2. Starting the Controller

The `bin` directory of the Sequoia distribution contains the scripts to start the controller. Unix users must start the controller with **controller.sh** whereas Windows users will use **controller.bat**.

Sequoia Controller startup is tuned via a configuration file, called `controller.xml`, included under the `config/controller` directory of your Sequoia installation. A simple configuration file looks like this:

A standard Sequoia Controller configuration file looks like this:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE SEQUOIA-CONTROLLER PUBLIC "-//Continuent//DTD SEQUOIA-CONTROLLER 2.10//EN"  "http://sequ
<SEQUOIA-CONTROLLER>
<Controller port="25322">
    <Report hideSensitiveData="true" generateOnFatal="true"/>
    <JmxSettings>
      <RmiJmxAdaptor/>
    </JmxSettings>
  </Controller>
</SEQUOIA-CONTROLLER>
```

You can specify at startup a different file than `config/controller/controller.xml`. This is useful if you have to startup many identical controllers from the network. You can then use the command **controller.sh -f filename** on Unix machines or **controller.bat -f filename** on windows.

For more information you can refer to the controller-configuration.xml example in the example directory of sequoia.

Next section describes how to write a controller configuration file.

7.3. Writing the controller configuration file

The controller is entirely configurable via an xml file, by default it is `controller.xml` located in the `config/controller` of the Sequoia installation. This section details how to write such a file.

7.3.1. Controller Parameters

The root element of the controller configuration is defined as follows

```
<!ELEMENT Controller (Internationalization?, Report?, JmxSettings?,
                                    VirtualDatabase*, SecuritySettings?)>
<!ATTLIST Controller
  port             CDATA "25322"
  ipAddress        CDATA "127.0.0.1"
  backlogSize      CDATA "10"
>
```

All sub-elements of Controller are defined in the next sections. Here is a brief overview of each of them:

- `Internationalization`: defines the language setting for Sequoia console and error messages.
- `Report`: if this option is enabled, Sequoia can automatically generate a report on fatal errors or shutdown. If you experience any problem with Sequoia, you can directly send the report on the mailing list to get a quick diagnostic of what happened.

- `JmxSettings`: JMX is the technology used for management and monitoring in Sequoia. These functionalities can be accessed through HTTP with an Internet browser or through the RMI connector used by the Sequoia console.
- `VirtualDatabase`: Defines a virtual database to load automatically at controller startup given a reference to its configuration file.
- `SecuritySettings`: Allows to filter accesses to a controller based on access lists.

The attributes of a Controller element are defined as follows:

- `port`: the port number on which clients (Sequoia drivers) will connect. The default port number is *25322*.

  Note: A port number below 1024 will require running the controller with privileged rights (root user under Unix).

- `ipAddress`: This can be defined to bind a specific IP address in case of a host with multiple IP addresses. This can be ignored if there is only one IP address available and will be replaced by `127.0.0.1`.
- `backlogSize`: the server socket backlog size (number of connections that can wait in the accept queue before the system returns "connection refused" to the client). Default is 10. Tune this value according to your operating system, but the default value should be fine for most settings.

If your machine has multiple network adapters, you can for the Sequoia Controller to bind a specific IP address like this:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE SEQUOIA-CONTROLLER PUBLIC "-//Continuent//DTD SEQUOIA-CONTROLLER 2.10//EN"  "http://seq
<SEQUOIA-CONTROLLER>
<Controller port="25322" ipAddress="192.168.0.1">
<JmxSettings enabled="false"/>
</Controller>
</SEQUOIA-CONTROLLER>
```

7.3.2. Internationalization

You can use this element to override the default locale retrieved by java. English is the only language looked at at the moment.

```
<!ELEMENT Internationalization EMPTY>
<!ATTLIST Internationalization language (en|fr|it|jp) "en">
```

7.3.3. Report

A report can be define in case you want to get a trace of what happened during the execution of the controller. If this element is included in the `controller.xml` report is enabled and will output a report, under certain conditions, in a file named `sequoia.report`.

```
<!ELEMENT Report EMPTY>
<!ATTLIST Report
```

```
      hideSensitiveData  (true|false) "true"
      generateOnShutdown (true|false) "true"
      generateOnFatal    (true|false) "true"
      enableFileLogging  (true|false) "true"
      reportLocation     CDATA        #IMPLIED
>
```

- `hideSensitiveData`: will replace passwords with '*****'.
- `generateOnShutdown`: tells the controller to generate a report when it has received a shutdown command.
- `generateOnFatal`: tells the controller to generate a report when it cannot recover from an error.
- `enableFileLogging`: logs all the console output into a file and include this file into the report.
- `reportLocation`: specify the path where to create the report, default is `SEQUOIA_HOME/log` directory.

7.3.4. JMX

JMX is used to remotely administrate the controller. You can use the bundled Sequoia console or use your own code to access JMX MBeans via the protocol adaptor. Sequoia proposes both the RMI and HTTP adaptors of the MX4J (http://mx4j.sourceforge.net/) JMX server. You can override the default port numbers for each adaptor if they conflict with another application that is already using them (i.e. another Sequoia controller on the same machine).

```
    <!ELEMENT JmxSettings (HttpJmxAdaptor?, RmiJmxAdaptor?)>
    <!ELEMENT HttpJmxAdaptor EMPTY>
    <!ATTLIST HttpJmxAdaptor
      port CDATA "8090"
    >

     <!ELEMENT RmiJmxAdaptor (SSL?)>
     <!ATTLIST RmiJmxAdaptor
       port         CDATA        "1090"
       username     CDATA        #IMPLIED
       password     CDATA        #IMPLIED
     >


     <!ELEMENT SSL EMPTY>
     <!ATTLIST SSL
       keyStore CDATA        #REQUIRED
       keyStorePassword CDATA        #REQUIRED
       keyStoreKeyPassword CDATA        #IMPLIED
       isClientAuthNeeded (true|false) "false"
       trustStore CDATA        #IMPLIED
       trustStorePassword CDATA        #IMPLIED
     >
```

Configure ssl for encryption and/or authentication.

- `keyStore`: The file where the keys are stored
- `keyStorePassword`: the password to the keyStore

- `keyStoreKeyPassword`: the password to the key, if none is specified the same password as for the store is used
- `isClientAuthNeeded`: if set to false ssl is used for encryption, the server is only accepting trusted clients (the client certificate has to be in the trusted store)
- `trustStore`: the file where the trusted certificates are stored, if none is specified the same store as for the key is used
- `trustStorePassword`: the password to the trustStore, if none is specified the same password as for the keyStore is used

You have to enable the RMI adaptor if you want to use the Sequoia console to administrate the controller remotely. To enable the RMI JMX adaptor, use this setting:

```
     <JmxSettings>
       <RmiJmxAdaptor/>
 </JmxSettings>
```

7.3.5. Virtual Database

This element specifies virtual databases to load at controller startup.

```
<!ELEMENT VirtualDatabase EMPTY>
<!ATTLIST VirtualDatabase
    configFile          CDATA #REQUIRED
    virtualDatabaseName CDATA #REQUIRED
    autoEnableBackends  (true | false | force) "true"
    checkpointName      CDATA ""
>
```

- `configFile`: The path to the virtual database configuration file. See Section 10 to learn how to write a virtual database configuration file.
- `virtualDatabaseName`: The name of the virtual database since the configuration file can contain multiple virtual database definitions.
- `autoEnableBackends`: set to true by default to reenable backends from their last known state as stored during last shutdown. If backends where not properly shutdown, nothing will happen. You can specify false to let the backends in disabled state at startup. The force option should only be used if you know exactly what you are doing and override backend status by providing a new checkpoint. *Warning!* Use this setting carefully as it might break your database consistency if you do not provide a valid checkpoint.Force is considered the same as true if no recovery log has been defined.
- `checkpointName`: the checkpoint name to use with the recovery log to enable backend from a known coherent state. If the checkpoint is omitted, the last known checkpoint is used.

Example:

```
<VirtualDatabase configFile="/databases/MySQLDb.xml" virtualDatabaseName="rubis" autoEnableBacken
```

This will enable a virtual database named `rubis` taken from a configuration file named `/databases/MySQLDb.xml` and will enable all backends of the database from the last known checkpoint.

7.3.6. Security

Security settings define the policy to adopt for some functionalities that may compromise the security of the controller. These settings depends on your environment and can be relaxed if you are running in a secure network. The less security settings you have, the faster the controller will run. A SecuritySettings element is defined as follows:

```
<!ELEMENT SecuritySettings (Jar?, Accept?, Block?, SSL?)>
<!ATTLIST SecuritySettings
  defaultConnect (true|false) "true"
>
```

`defaultConnect`: is used to allow (true) or refuse (false) connections to the controller. This default setting can be then be tuned with access lists defined in `Accept` and `Block` elements (see below).

Additional database drivers can be uploaded dynamically to the controller. As the controller has no way to check if this is a real JDBC driver or some malicious code hidden a JDBC driver interface, you have to be very careful if you enable this option and anybody can connect from anywhere to your controller.

```
<!ELEMENT Jar EMPTY>
<!ATTLIST Jar
allowAdditionalDriver (true|false) "true"
>
```

You can control who can connect to the controller by setting access lists based on IP addresses to accept or block. `defaultConnect` is set in `SecuritySettings` defined above. Default is to accept all connections if no security manager is enabled.

```
<!ELEMENT Accept (Hostname|IpAddress|IpRange)*>
<!ELEMENT Block (Hostname|IpAddress|IpRange)*>

<!ELEMENT Hostname EMPTY>
<!ATTLIST Hostname
     value CDATA #REQUIRED
>
```

IpAddress value is an IPv4 address (ex:192.168.1.12):

```
<!ELEMENT IpAddress EMPTY>
<!ATTLIST IpAddress
     value CDATA #REQUIRED
>
```

IpRange value is based on IPv4 addresses and has the following form: 192.168.1.*.

```
<!ELEMENT IpRange EMPTY>
<!ATTLIST IpRange
     value CDATA #REQUIRED
>
```

Here is a full security configuration example:

```
<SecuritySettings defaultConnect="false">
```

```
  <Jar allowAdditionalDriver="true"/>
  <Shutdown>
    <Client allow="true" onlyLocalhost="true"/>
    <Console allow="false"/>
  </Shutdown>
  <Accept>
    <IpRange value="192.168.*.*"/>
  </Accept>
</SecuritySettings>
```

This setting accepts driver connections only from machines having an IP address starting with 192.168, allows loading of additional drivers via the console, refuses shutdown from the console, but allows it from the local machine.

7.4. Configuring the Log

Sequoia uses the Log4j (http://jakarta.apache.org/log4j/) logging framework. The `log4j.properties` configuration file is located in the `/sequoia/config` directory of your installation. Here is a brief description of the loggers available in the configuration file:

- `log4j.logger.org.continuent.sequoia.core.controller` : Controller related activities mainly for bootstrap and virtual database adding/removal operations.
- `log4j.logger.org.continuent.sequoia.controller.xml.Handler` : XML configuration file parsing and handling.
- `log4j.logger.org.continuent.sequoia.controller.VirtualDatabase` : Virtual database related operations. A specific `log4j.logger.org.continuent.sequoia.controller.VirtualDatabase.virtualDatabaseName` logger is automatically created for each virtual database. This allows to tune different logging levels for each virtual database.
- `log4j.logger.org.continuent.sequoia.controller.VirtualDatabase.request` : Log the incoming requests and transactions in files that can be replayed by the Request Player tool provided with Sequoia.
- `log4j.logger.org.continuent.sequoia.controller.distributedvirtualdatabase.request` : Log distributed request execution when using horizontal scalability (a.k.a. controller replication).
- `log4j.logger.org.continuent.sequoia.controller.backup` : Log backup manager and backuper related activities from dump/restore operations.
- `log4j.logger.org.continuent.sequoia.controller.VirtualDatabaseServerThread` : The server thread accepts client connections and manages the worker threads.
- `log4j.logger.org.continuent.sequoia.controller.VirtualDatabaseWorkerThread` : Each worker thread handle a session with a client Sequoia driver.
- `log4j.logger.org.continuent.sequoia.controller.RequestManager` : Log the request flows between the different Request Manager components (scheduler, cache, load balancer, recovery log).
- `log4j.logger.org.continuent.sequoia.controller.scheduler` : Log the request ordering and synchronization performed by the scheduler.
- `log4j.logger.org.continuent.sequoia.controller.cache` : SQL Query cache related activities.

- `log4j.logger.org.continuent.sequoia.controller.loadbalancer` : Log how requests are balanced on the backends.
- `log4j.logger.org.continuent.sequoia.controller.connection` : Connection pooling related information.
- `log4j.logger.org.continuent.sequoia.controller.recoverylog` : Sequoia Recovery Log information.
- `log4j.logger.org.continuent.sequoia.controller.console.jmx` : JMX management system logging.
- `log4j.logger.org.continuent.hedera.channels`: Hedera low level group communication channel.
- `log4j.logger.org.continuent.hedera.gms`: Hedera Group Membership Service (GMS).
- `log4j.logger.org.continuent.tribe.discovery`: Tribe Discovery Service (used by GMS).
- `log4j.logger.org.continuent.hedera.adapters`: Hedera Multicast Dispatcher building block for application level message handling.
- `log4j.logger.org.jgroups`: JGroups core messages when Hedera is used with JGroups.
- `log4j.logger.org.jgroups.protocols`: JGroups protocol stack messages when Hedera is used with JGroups.

7.5. Recovery Log

When you want to add a database to your cluster, you do not want to stop the system, replicate the current database state to the new database (that may take a long while) and then restart the system. The Recovery Log helps you in the process of dynamically adding a new backend (or recovering a previously failed backend) without stopping the system.

The Recovery Log records the write operations and transactions that are performed by the Sequoia controller between checkpoints. A checkpoint is just a logical index in the log that reflect the recovery log state at a given time. As of Sequoia 2.0, checkpoints are automatically managed by the controller and are generated when needed on behalf of the administrator when a backend is disabled or enter a backup phase. When re-enabling the backend, the Recovery Log replays all write queries and transactions that the backend missed during the time it was offline and it comes back to the enabled state once it is synchronized with the other nodes.

Since version 2.0, the backup infrastructure has completely changed and is based on Backupers. We provide a generic Backuper based on Enhydra Octopus (http://octopus.enhydra.org/) to copy, backup and restore content of backends through JDBC. Even if Octopus is supposed to handle most common databases, it might fail for some specific databases or data types. In that case, we strongly recommend to use or implement a database specific Backuper.

> **Note:** Octopus currently fails to backup/restore empty databases. You need at least to have one table in your database if you don't want the backup operation to fail with Octopus.

7.5.1. A practical example

Your Web site is running with a single database and you want to use Sequoia with three nodes using full replication (RAIDb-1). You have two new backends ready to be installed. You can start the Sequoia console and connect to the controller. Start the administration module by connecting to the virtual database. Type: **backup <backend name> <dump name> <backuper name> <path to backup directory>**. If you want to use Octopus you will use a command line like **backup node1 dump1 Octopus /var/backups**. During the backup, the update

requests are logged in the recovery log, so no update is lost. If the backend was in the enabled state when backup was initiated, it will automatically replay the recovery log to resynchronize itself and return to the enabled state.

To restore the dump on another backend, just type **restore <newbackend> <dumpname>** and the appropriate backuper (Octopus in our previous example) will be used to restore the dump. After restoring the dump, you can enable the backend at any time so that the recovery log replays all the missing requests since the dump was taken.

Here is the set of commands to use in the Sequoia console if node1 is your existing backend and you want to dynamically add node2 and node3:

```
backup node1 initial_dump Octopus /var/backups
restore node2 initial_dump
restore node3 initial_dump
enable node2
enable node3
```

Note: Note that these steps can be automated by scripting the console.

If a node crashes, use the administration console to restore the dump on the node using the restore command. Once the dump is restored, re-enable the backend from the stored checkpoint and the Recovery Log will automatically replay all the write queries to rebuild a consistent database state on the node.

To prevent the recovery log from being too large, you can periodically perform backup operations. This will also lower the recovery time since the part of the log to replay will be smaller. You can delete older dumps and logs if you do not need them anymore.

7.5.2. Understanding checkpoints

A checkpoint is a reference used by the recovery log to replay missing requests. If a backend is disabled from the console for maintenance, the controller will automatically create a checkpoint (in C-JDBC, the checkpoint name had to be provided manually through the console). Once the backend is enabled again, the controller retrieves its last known checkpoint from the recovery log and replays all the requests that the disabled backend missed since it was disabled. A checkpoint is nothing more than a reference in time.

7.5.3. A fault tolerant Recovery Log

As the Sequoia recovery log can be stored in a database providing a JDBC driver, it is possible to make the recovery log fault tolerant by redirecting it to a Sequoia controller (even self) that will distribute and replicate the log content on several backends.

The JDBC Recovery Log configuration is detailed in Section 10.6.5.

7.6. Controller replication

To prevent the Sequoia controller from being a single point of failure, Sequoia provides controller replication also called horizontal scalability. A virtual database can be replicated in several controllers that can be added dynamically at runtime. Controllers use the JGroups group communication middleware to synchronize updates in a distributed way. The JGroups stack configuration is found in `config/jgroups.xml` and should not be altered unless you specifically know what you are doing. Keep in mind that total order reliable multicast is needed to

ensure proper synchronization of the controllers. More information about JGroups can be found on the JGroups web site (http://www.jgroups.org). Note that JGroups requires proper network settings, here are a few guidelines:

- a default route must be defined (check with /sbin/route under Linux) for the network adapter which is bound by JGroups (usually eth0). If such route does not exist, either the group communication initialization will block or controllers will not be able to see each other even on the local host. If you don't have any default entry in your routing table you can use a command like '/sbin/route add default eth0' to define this default route.
- issues have been reported with DHCP that can either block (under Windows) or just fail to properly set a default route and leads to the issue reported above. We strongly discourage the use of DHCP, you should use fixed IP addresses instead.
- name resolution should be properly set so that the IP address/machine name matching works both ways. Often improper /etc/hosts or DNS configuration leads to group communication initialization problems. In particular, under Linux, the IP address associated to the name returned by the 'hostname' command must not resolve to 127.0.0.1 else controllers will not see each other.

Horizontal scalability can also be provided using Appia. The Appia stack configurations are found in `config/appia.xml`. This file contains six different configurations, six templates for communication channels and their respective channel instantiations. These are the combinations of two total order implementations (sequencer based and token based total order) using different transport protocols: TCP, UDP and UDP multicast. Instructions to change the default configuration are in the header of the file. All the defined configurations ensure total order reliable multicast. More information about Appia can be found on the web site (http://appia.continuent.org). Note that Appia also requires proper network settings, here are a some guidelines:

- a default route must be defined (check with /sbin/route under Linux) for the network adapter which is bound by Appia (usually eth0). If such route does not exist, controllers will not be able to see each other. If you don't have any default entry in your routing table you can use a command like '/sbin/route add default eth0' to define this default route.
- name resolution should be properly set so that the IP address/machine name matching works both ways. Often improper /etc/hosts or DNS configuration leads to group communication initialization problems. In particular, under Linux, the IP address associated to the name returned by the 'hostname' command must not resolve to 127.0.0.1 else controllers will not see each other.
- Appia does not need to use fixed IP addresses, unless you want to bind a controller to a specific IP address. To discover other controllers Appia uses a gossip service. The gossip service can be configured to use a multicast address (if your network supports it) or you can start a gossip server. This server can also be replicated and is used just to help the dynamic discovery of new nodes.

In order for a virtual database to be replicated, you must define a `Distribution` element in the virtual database configuration file (see Section 10.2.1). There are several constraints for different controllers to replicate a virtual database:

- give the list of all controllers that you plan to use for replication of your virtual database in the Sequoia driver URL. Even if all controllers are not online at all times, the driver will automatically detect the alive controllers: jdbc:sequoia://node1,node2,node3,node4/myDB
- the virtual database must have the same name and use the same groupName (in the Distribution element).
- each controller must have its own set of backends and no backends should be shared between controllers (Sequoia checks the database URLs, having different backend names is not sufficient).
- each controller must have its own recovery log, recovery logs cannot be shared. It is possible for a controller not to have a recovery log but this controller will have no recovery capabilities.
- the authentication managers must support the same logins.

- schedulers and load balancers must implement the same RAIDb configuration.
- database schemas (if defined) must be compatible according to the RAIDb level you are using.

> **Note:** As backends cannot be shared between controllers, it is not possible to use a SingleDB load balancer with controller replication. If each controller only has a single database backend attached to it, then you must use a RAIDb-1 configuration since in fact you have 2 replicated backends in the cluster.

Several configuration file examples are available in the `doc/examples/HorizontalScalability` directory of your Sequoia distribution.

> **Note:** You can find more information in the document titled "Sequoia Horizontal Scalability - A controller replication user guide" available from the Sequoia web site.

7.7. Current Limitations

The Sequoia controller in its 2.10 release has the following limitations:

- GRANT/REVOKE commands will be sent to the database engines but this will not add or remove users from the virtual database authentication manager.
- network partition/reconciliation is not supported,
- distributed joins are not supported which means that you must ensure that every query can be executed by at least a single backend,
- `RAIDb-1ec` and `RAIDb-2ec` levels are not supported,

8. Administration console

The Sequoia administration console is based on JMX technologies. The text mode console is a JMX client based on the standard RMI connector for JMX but you can also use a generic a JMX administration console through HTTP from any web browser to see all the MBeans registered in the sequoia domain. An Eclipse plug-in is provided by the Oak project (http://oak.continuent.org).

You can start the administration console using the console.sh/.bat script.

8.1. Jmx Notifications List

Here is a list of the JMX remote notifications generated by Sequoia.

- `sequoia.controller.virtualdatabases.removed` a virtual database has been removed.
- `sequoia.controller.virtualdatabase.added` a virtual database has been added to the controller
- `sequoia.virtualdatabase.dump.list` the list of dump files has been updated
- `sequoia.virtualdatabase.backend.added` a backend has been added to the virtual database
- `sequoia.distributed.controller.added` a controller has joined the group

- `sequoia.virtualdatabase.backend.disabled` a backend has been disabled
- `sequoia.virtualdatabase.backend.enabled` a backend has been enabled
- `sequoia.virtualdatabase.backend.recovering` a backend is recovering a dump file
- `sequoia.virtualdatabase.backend.recovery.failed` Recovery of a dump file failed
- `sequoia.virtualdatabase.backend.replaying.failed` Recovery log replay failed
- `sequoia.virtualdatabase.backend.backingup` a backend is backing up
- `sequoia.virtualdatabase.backend.enable.write` a backend is now write enabled
- `sequoia.virtualdatabase.backend.removed` a backend has been removed from the virtual database
- `sequoia.virtualdatabase.backend.disabling` a backend is now in state disabling (finishing pending transactions and pending requests)
- `sequoia.virtualdatabase.backend.unknown` The backend state has been completely lost. Recovery needed
- `sequoia.virtualdatabase.backend.replaying` a backend is replaying requests from the recovery log

8.2. Starting the Administration Console

The `bin` directory of the Sequoia distribution contains the scripts to start the console. Unix users must start the console with **console.sh -t** whereas Windows users have to start **console.bat -t**.

The console script accepts several options:

- `-d` or `--debug`: show stack trace when error occurs.
- `-f` or `--file`: Use a given file as the source of commands instead of reading commands interactively.
- `-h` or `--help`: displays usage information.
- `-i` or `--ip`: IP address of the host name where the JMX Server hosting the controller is running (the default is '0.0.0.0').
- `-p` or `--port`: JMX/RMI port number of (the default is 1090).
- `-s` or `--secret`: Password for JMX connection.
- `-u` or `--username`: username for JMX connection.
- `-v` or `--version`: displays version information.
- `-t` or `--text`: force the console to start in text mode. By default, it will try to start in graphic mode

For example, **console.sh -t -i 192.168.0.1 -p 1234** will connect the console to the controller using the RMI JMX adaptor listening on port 1234 on `192.168.0.1`.

The console has an online help that is accessible by typing **help** at any time.

8.3. Console Quickstart

Here is a quick description of the steps needed to make a controller ready to serve requests:

1. Start the controller using **controller.sh** or **controller.bat** (see Section 7.2).
2. Start the console using **console.sh -t** or **console.bat -t** (see Section 8.2).

3. Load a configuration file using **load <complete-path>/config.xml**. The controller configuration files are described in Section 10.
4. Connect to the virtual database with the administrator login using the **admin** command (see example below).
5. Enable all backends using the **enableAll** command.
6. Come back to the main menu using the **quit** command.
7. Check the configuration using the **getInfo** command.

Here is an example of a controller configuration and startup:

```
[emmanuel@gre-home bin]$ console.sh -t
Launching the Sequoia controller console
Initializing Controller module...
Initializing VirtualDatabase Administration module...
Initializing Monitoring module...
Initializing SQL Console module...
Sequoia driver (v. 2.0) successfully loaded.

gre-home:1090 >help
Commands available for the Controller module are:
admin <virtualdatabase name>
   Administrate a virtual database
connect controller <controller hostname> <jmx port>
   Connect to a Sequoia controller
drop virtualdatabase <virtualdatabase name>
   Drop a virtual database from the controller
help
   Print this help message
history [<commandIndex>]
   Display history of commands for this module
load virtualdatabase config <virtualdatabase xml file>
   Send a virtual database XML configuration file to the controller and load it
monitor <virtualdatabase name>
   Monitor a virtual database
quit
   Quit this console
reload logging configuration
   Refresh the trace system by reloading the logging configuration file
save configuration
   Save the current configuration of the virtual databases as an XML file
show controller config
   Show Controller configuration
show logging config
   Show logging configuration and the most recent traces
show virtualdatabases
   Show the names of the virtual databases for this controller
shutdown [mode]
   Shutdown the controller and all its virtual databases. Mode parameter must be:
        1 -- wait for all client connections to be closed, does not work with a connection pool
        2 -- mode safe, default value, waits for all current transactions to complete
        3 -- mode force, immediate shutdown without consistency: recovery will be needed on resta
sql client <sequoia url>
   Open a SQL client console for the virtual database specified by the Sequoia URL
upload driver <driver file>
   Upload a driver to the controller
```

```
gre-home:1090 > <userinput>show virtualdatabases</userinput>
myDB
gre-home:1090 > <userinput>admin myDB</userinput>
Virtual database Administrator Login > <userinput>admin</userinput>
Virtual database Administrator Password > <userinput>*****</userinput>
Ready to administrate virtual database myDB
myDB(admin) > help
Commands available for the VirtualDatabase Administration module are:
backup <backend name> <dump name> <backuper name> <path> [<tables>]
   Backup a backend into a dump file and associate a checkpoint with this dump
delete dump <dump name>
   Delete a dump
disable <backend name | *>
   Disable the specified backend and automatically set a checkpoint
        * means that all backends of this virtual database must be disabled
enable <backend name | *>
   Enable the specified backend
        * means that all backends of this virtual database must be enabled
expert <on|off>
   Switch to expert mode (commands for advanced users are available)
help
   Print this help message
history [<commandIndex>]
   Display history of commands for this module
quit
   Quit this console
restore <backend name> <dump name> [<tables>]
   Starts the recovery process of the given backend for a given dump name
show backend <backend name | *>
   Show information on backend of this virtual database
        * means to show information for all the backends of this virtual database
show backends
   Show the names of the backends of this virtual database on the current controller
show backupers
   Show the backupers available for backup
show controllers
   Show the names of the controllers hosting this virtual database
show dumps
   Show all dumps available for database recovery
show virtualdatabase config
   Show the XML configuration of the virtual database
transfer dump <dump name> <controller name> [nocopy]
   Make a dump available for restore on another controller.
Optional 'nocopy' (default: false) flag specifies not to copy the dump.

myDB(admin) > <userinput>show backend *</userinput>
+-------------------------+----------------------------------+
| Backend Name            | localhost                        |
| Driver                  | org.hsqldb.jdbcDriver            |
| URL                     | jdbc:hsqldb:hsql://localhost:9001 |
| Active transactions     | 0                                |
| Pending Requests        | 0                                |
| Read Enabled            | true                             |
| Write Enabled           | true                             |
| Is Initialized          | true                             |
| Static Schema           | false                            |
| Connection Managers     | 1                                |
```

```
| Total Active Connections | 5                                  |
| Total Requests           | 0                                  |
| Total Transactions       | 0                                  |
| Last known checkpoint    | <unknown>                          |
+--------------------------+------------------------------------+
| Backend Name             | localhost2                         |
| Driver                   | org.hsqldb.jdbcDriver              |
| URL                      | jdbc:hsqldb:hsql://localhost:9002 |
| Active transactions      | 0                                  |
| Pending Requests         | 0                                  |
| Read Enabled             | true                               |
| Write Enabled            | true                               |
| Is Initialized           | true                               |
| Static Schema            | false                              |
| Connection Managers      | 1                                  |
| Total Active Connections | 5                                  |
| Total Requests           | 0                                  |
| Total Transactions       | 0                                  |
| Last known checkpoint    | <unknown>                          |
+--------------------------+------------------------------------+
```

8.4. Console Main Menu

The graphical version of the console provides a shell-like history (more precisely a tcsh-like behavior). You can recall a previous command by using the arrow keys (up and down) to browse the history. If you prefix a command by `!`, the console will browse the history and complete the command with the latest command in the history starting with the command prefix (completion occurs when you press the tab key). In the graphical version, you can also access all the commands of the different module using the right button of the mouse.

> **Note:** All the commands issued can also be recalled using the history menu in the contextual menu that appears on a right-button click.

Commands available from the console main menu are:

- **admin <virtualdatabase name>**: Administrate a virtual database
- **connect controller <controller hostname> <jmx port>**: connect to a Sequoia controller
- **drop virtualdatabase <virtualdatabase name>**: Drop a virtual database from the controller
- **help**: Print this help message
- **history [<commandIndex>]**: Display history of commands for this module
- **load virtualdatabase config <virtualdatabase xml file>**: Send a virtual database XML configuration file to the controller and load it
- **monitor <virtualdatabase name>**: Monitor a virtual database
- **quit**: Quit this console
- **reload logging configuration**: Refresh the trace system by reloading the logging configuration file
- **save configuration**: Save the current configuration of the virtual databases as an XML file

- **show controller config**: Show Controller configuration
- **show logging config**: Show logging configuration and the most recent traces
- **show virtualdatabases**: Show the names of the virtual databases for this controller
- **shutdown [mode]**: shutdown the controller and all its virtual databases.

  Three shutdown modes are provided. If not specified, the default mode is the shutdown mode immediate.

  - Shutdown mode wait (mode `1`): wait for all client connections to be closed, does not work if the client uses a connection pool with persistent connections.
  - Shutdown mode safe (mode `2`): default value, waits for all current transactions to complete before shutting down. transaction and shutdown.
  - Shutdown mode force (mode `3`): does not wait for transactions completion and kill all connections. Backends are disabled without consistency and a full recovery will be needed on restart.

  E.g: **shutdown 2**.

- **sql client <sequoia url>**: Open a SQL client console for the virtual database specified by the Sequoia URL
- **upload driver <driver file>**: Upload a driver to the controller

8.5. Administrator Menu

Once the configuration file has been loaded on the controller, all backends are in the disabled state. You must enable them all or one by one to allow them to execute requests. Sequoia does not check that database contents are synchronized and you must ensure that all backends are in a coherent state prior to starting the controller. To ensure that backends remain synchronized on startup, you must use checkpoints (see Section 7.5.2).

If you properly shutdown the controller using the wait or safe mode, database backend states are properly recorded and their state is automatically restored when they are enabled.

8.5.1. Administrator Standard Commands

Standard commands available from the console administrator menu are:

- **backup <backend name> <dump name> <backuper name> <path> [<tables>]**: Backup a backend into a dump file and associate a checkpoint with this dump. Note that the console will ask for a login and password to connect to the backend to backup. This is specific to the Backuper that you are using but this should usually be a valid login/password on the database engine that you are backuping. The login must be granted access on all tables from the controller node.
- **delete dump <dump name>**: Delete a dump
- **disable <backend name | *> <checkpoint>**: Disable the specified backend and store the given checkpoint (* means that all backends of this virtual database must be disabled)
- **enable <backend name | *>**: Enable the specified backend from its last known checkpoint (* means that all backends of this virtual database must be enabled)
- **expert <on|off>**: Switch to expert mode (commands for advanced users are available)
- **help**: Print this help message
- **history [<commandIndex>]**: Display history of commands for this module

- **quit**: Quit this console
- **restore <backend name> <dump name> [<tables>]**: Starts the recovery process of the given backend using the given dump name. Note that the console will ask for a login and password to connect to the backend to restore This is specific to the Backuper that you are using but this should usually be a valid login/password (real login in the Sequoia terminology) on the database engine that you are restoring. Note that this login must be granted the right to create new databases and tables.
- **show backend <backend name | *>**: Show information on backend of this virtual database (* means to show information for all the backends of this virtual database)
- **show backends**: Show the names of the backends of this virtual database on the current controller
- **show backupers**: Show the backupers available for backup
- **show controllers**: Show the names of the controllers hosting this virtual database
- **show dumps**: Show all dumps available for database recovery
- **show virtualdatabase config**: Show the XML configuration of the virtual database
- **transfer dump <dump name> <controller name>**: Transfer a dump from the current controller to another controller. An example is **transfer dump dump1 controller2.emic.com:1090**

8.5.2. Administrator Expert Commands

Expert commands are not available by default. use the command expert on to make them available.

- **clone backend config <backend from> <backend to> <url> [driverPath=<value>] [driver=<value>] [connectionTestStatement=<value>]**: Clone the configuration of a backend in the current virtual database (this virtually allows to add a new backend)
- **disable read <backend name>**: Disable read requests on a backend
- **enable read <backend name>**: Enable read requests on a backend
- **force checkpoint <backend name> <checkpoint name>**: Force the last know checkpoint of a disabled backend
- **force disable <backend name | *>**: Force the disabling of a backend without storing any checkpoints. The backend will not be in a consistent state after this operation! (* means that all backends of this virtual database must be disabled by force)
- **force enable <backend name | *>**: Force the enabling of a backend without checking for checkpoints. This command can break the cluster consistency, only use it if you know what you are doing! (* means that all backends of this virtual database must be enabled by force)
- **force path <dump name> <new path>**: Update the path of the dump
- **get backend schema <backend name> <file name>**: Display backend schema or save it to a file
- **purge log <checkpoint name>**: Purge the recovery log upto specified checkpoint. All the entries of the recovery log prior to that checkpoint will be deleted.
- **restore log <dump name> <controller name>**: Copy the local recovery log from the specified checkpoint onto the specified remote controller. All previous recovery log content on the remote controller will be erased.
- **show checkpoints**: Show all checkpoints available in the recovery log.
- **transfer backend <backend name> <controller jmx address>**: Transfer a backend from a controller to an other controller

8.6. Automated Backup With Jmx

Marc Wick has given an example of a cron file to do a daily backup using the jmx connector in Sequoia. The complete sources can be found in the example file:DBBackup.java in the jmx directory of the examples.

```
JMXServiceURL address = new JMXServiceURL("rmi", host, 0, "/jndi/jrmp");

Map environment = new HashMap();
environment.put(Context.INITIAL_CONTEXT_FACTORY,
    "com.sun.jndi.rmi.registry.RegistryContextFactory");
environment.put(Context.PROVIDER_URL, "rmi://" + host + ":" + port);
environment.put(JMXConnector.CREDENTIALS, PasswordAuthenticator
    .createCredentials("jmxuser", "jmxpassword"));

JMXConnector connector = JMXConnectorFactory.connect(address, environment);
ObjectName db = JmxConstants.getVirtualDbObjectName("databaseName");

...

MBeanServerConnection delegateConnection = connector
    .getMBeanServerConnection(subj);

// we create a proxy to the virtual database
VirtualDatabaseMBean proxy = (VirtualDatabaseMBean) MBeanServerInvocationHandler
    .newProxyInstance(delegateConnection, db, VirtualDatabaseMBean.class,
        false);

SimpleDateFormat fmt = new SimpleDateFormat("yyyy_MM_dd");
String checkpointName = fmt.format(new Date());

// we disable the backend and set a checkpoint
proxy.disableBackendForCheckpoint("node1", checkpointName);

// we call the database specific backup tool for the backup
runDatabaseBackupTool();

// we enable the backend again
proxy.enableBackend("node1");
```

The runDatabaseBackupTool() method is completely open and can call any external program (like pg_dump, mysql_dump...)

> **Note:** This method does not use octopus and as a consequence, the generated backup cannot be restored on a a different database vendor than the one it was issued from. As a great benefit though, the backup process will gain in speed, and the metadata will be completely conformed to that database vendor.

8.7. Recoverying from a failed controller in distributed mode

In a distributed controller configuration, when a controller goes down, here is the list of action to take to recover the failed controller:

- If the controller does not have any dump available, connect to a controller that has database dumps and use the **transfer dump** command to copy the dump to the recovering controller.

- During its failure, the recovery log of the controller missed queries that were executed by the cluster and it is therefore necessary to re-synchronize its recovery log. This can be achieve using the **recover log** from the same controller you used to transfer the dump.
- Once the previous operations are completed, you can safely restore the dump on the backends attached to the controller. Then, enabling the backends will resynchronize them with the other nodes of the cluster.

8.8. Virtual Database Console Menu

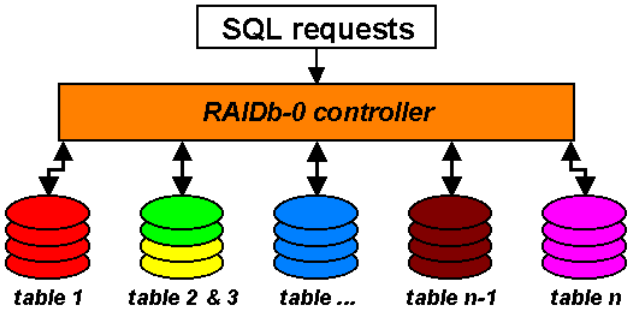
Sequoia is now bundled with a graphical SQL console called Squirrel that you can launch from the `bin` directory of the Sequoia installation, using either `squirrel.sh` or `squirrel.bat`.You can also directly issue SQL statements from the virtual database console menu. The other commands available from the virtual database console menu are:

- **begin**: Start a transaction
- **commit**: Commit a transaction
- **fetchsize <x>**: Set the ResultSet fetch size to x rows per block
- **help**: Print this help message
- **history [<commandIndex>]**: Display history of commands for this module
- **load <file name>**: Execute all SQL statements contained in file
- **maxrows <x>**: Limits the maximum number of rows to get from the database to x
- **quit**: Quit this console
- **rollback [<savepoint name>]**: Rollback a transaction (to an optional savepoint)
- **savepoint <savepoint name>**: Create a savepoint for the current transaction
- **setisolation <x>**: Set the connection transaction isolation level to x
  - 0 - TRANSACTION_NONE
  - 1 - TRANSACTION_READ_UNCOMMITTED
  - 2 - TRANSACTION_READ_COMMITTED
  - 4 - TRANSACTION_REPEATABLE_READ
  - 8 - TRANSACTION_SERIALIZABLE
- **show tables**: Display all the tables of this virtual database
- **timeout <x>**: Set the query timeout to x seconds (default is 60 seconds)
- **{call proc_name(?,?,...)}**: Call a stored procedure

Here is an example of a session with the virtual database console:

```
localhost:1090 > sql client jdbc:sequoia://localhost/myDB
 > Login     : user
 > Password  : ******
Connected to jdbc:sequoia://localhost/myDB
jdbc:sequoia://localhost/myDB (user) > begin
Transaction started
jdbc:sequoia://localhost/myDB (user) > select * from regions
... result to be displayed here ...
```

Figure 3. RAIDb-0 example




```
jdbc:sequoia://localhost/myDB (user) > commit
jdbc:sequoia://localhost/myDB (user) > quit
```

9. RAIDb Basics

9.1. RAIDb Definition

RAIDb stands for *Redundant Array of Inexpensive Databases*. This acronym has been used in reference to the RAID (*Redundant Array of Inexpensive Disks*) concept that achieves scalability and high availability of disk subsystems at a low cost. RAIDb aims at providing better performance and fault tolerance than a single database by combining multiple inexpensive database instances into an array of databases.

One of the goals of RAIDb is to hide the distribution complexity and to provide the database clients with the view of a single database. As for RAID, a controller sits in front of the underlying resources. The clients send their requests to the RAIDb controller that balances them among the set of RDBMS backends.

9.2. RAIDb-0

RAIDb-0 consists in *partitioning* the database tables among the database backend nodes. A table itself cannot be partitioned but the different tables can be distributed on different backend nodes. RAIDb-0 requires at least two database backends, provides moderate performance scalability but does not offer fault tolerance. Figure 3 shows an example of a RAIDb-0 configuration.

9.3. RAIDb-1

RAIDb-1 offers a *full mirroring* or *full replication* of the database on the backends. It offers the best fault tolerance scheme since the system is still available with only one backend. On the minus side, there is no speedup on writes (`UPDATE`, `INSERT`, `DELETE` requests) since they have to be broadcasted to all nodes. Figure 4 shows an example of a RAIDb-1 configuration.

Figure 4. RAIDb-1 example

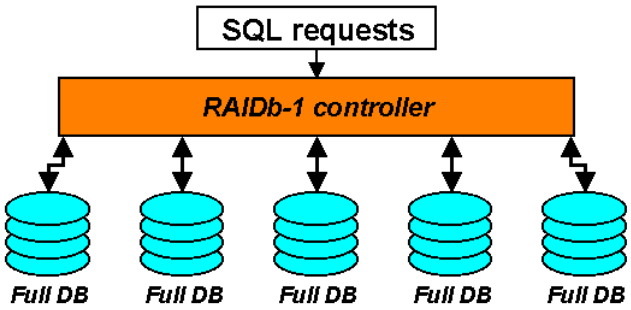
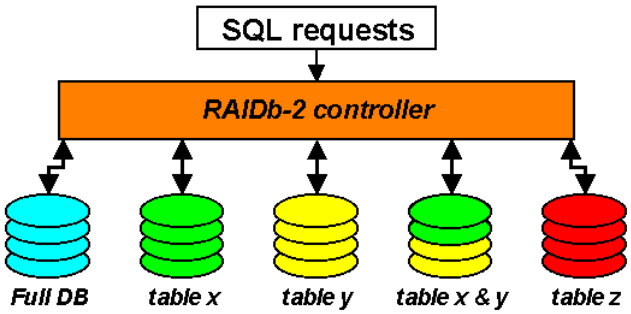


Figure 5. RAIDb-2 example



9.4. RAIDb-2

RAIDb-2 is a tradeoff between RAIDb-0 and RAIDb-1. It provides partial replication to tune the degree of replication of each database table to obtain the best read/write throughput. RAIDb-2 requires that each database table is available on at least two nodes. Figure 5 shows an example of a RAIDb-2 configuration.

9.5. Nested RAIDb Levels

It is possible to compose several RAIDb levels to build large scale configurations or meet specific needs. The next example is a RAIDb-1-0 configuration where a top level RAIDb-1 controller dispatches the requests to three full databases implemented with a RAIDb-0 controller. Figure 6 shows an example of a RAIDb-1-0 configuration.

This last example (Figure 7) shows a RAIDb-0-1 composition. The top level is a RAIDb-0 controller and fault tolerance is achieved on each partition using a RAIDb-1 controller.

Figure 6. RAIDb-1-0 example

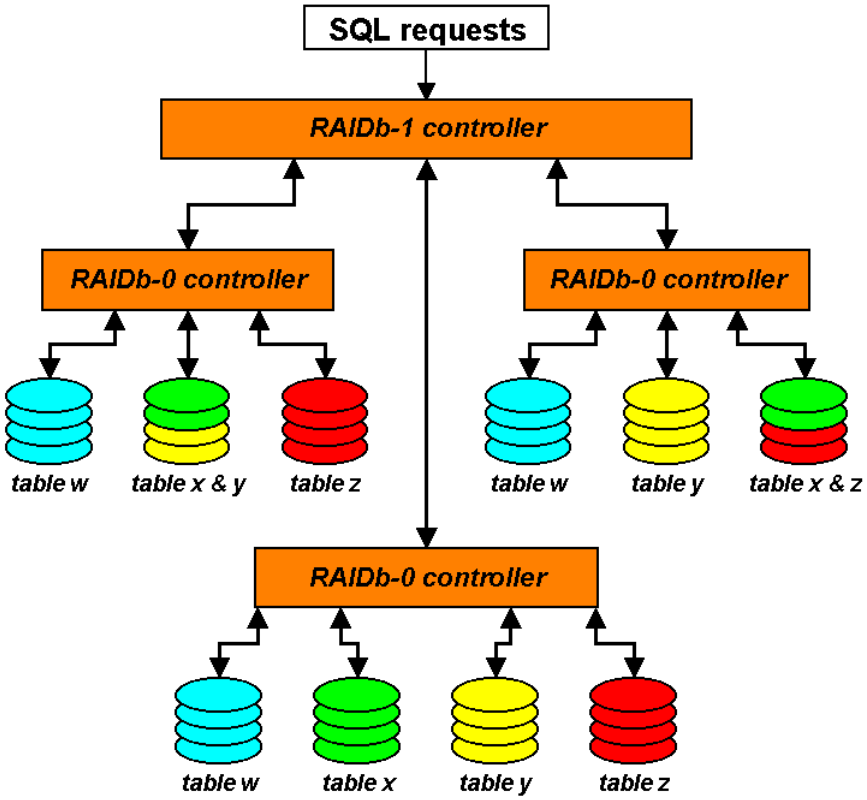
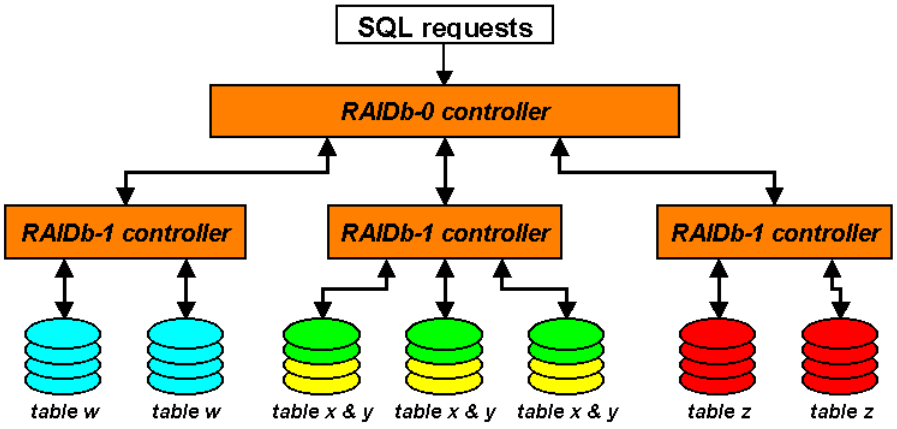


Figure 7. RAIDb-0-1 example



10. Virtual database configuration

10.1. Writing a Virtual Database Configuration File

The Sequoia controller configuration file must conform to the Sequoia DTD that can be found in the `xml` directory of the Sequoia distribution. The DTD is extensively documented and the most up-to-date information will be found in the `xml/sequoia-x.y.dtd` file. Several configuration file examples are available in the `doc/examples` directory.

Here is an example of how a minimal Sequoia configuration file should look like:

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE SEQUOIA PUBLIC "-//Continuent//DTD SEQUOIA 1.0//EN"
                         "http://sequoia.continuent.org/dtds/sequoia-2.10.dtd">

<SEQUOIA>
  <VirtualDatabase name="vdbName">

    <Distribution/>

    <AuthenticationManager> ...  </AuthenticationManager>

    <DatabaseBackend name="node1" driver="com.myDriver.class"
                     url="jdbc:protocol://host/myDB" connectionTestStatement="select 1">
       ...
    </DatabaseBackend>

    <RequestManager>
      <RequestScheduler>
         ...
      </RequestScheduler>

      <LoadBalancer>
         ...
      </LoadBalancer>
    </RequestManager>
  </VirtualDatabase>
</SEQUOIA>
```

The next sections describes the different elements composing an XML configuration file.

10.2. Virtual Database

A virtual database element is defined as follows:

```
<!ELEMENT VirtualDatabase (Distribution?, Monitoring?, Backup?,AuthenticationManager, DatabaseBacl
<!ATTLIST VirtualDatabase
       name               CDATA #REQUIRED
       maxNbOfConnections CDATA #IMPLIED
       poolThreads        (true | false) "true"
       minNbOfThreads     CDATA #IMPLIED
       maxNbOfThreads     CDATA #IMPLIED
       maxThreadIdleTime  CDATA #IMPLIED
```

```
        sqlDumpLength      CDATA "40"
        useStaticResultSetMetaData CDATA "true"
>
```

A virtual database is the database exposed to the user. It contains:

- a set of real database backends,
- an authentication manager that matches the virtual database and real backends login/password,
- a request manager that defines the behavior of the controller for this virtual database,

Here is a brief description of each virtual database attribute:

- `name`: name of the virtual database to be used in the JDBC URL (`jdbc:sequoia://host/VirtualDatabaseName`).
- `maxNbOfConnections`: maximum number of concurrent connections accepted for this virtual database. The controller stops accepting client connections when `maxNbOfConnections` concurrent connections are running. Default is 0 (no limit).
- `poolThreads`: if `false`, one thread is created for each connection and dies when the connection closes. If set to `true`, threads are created on-demand and kept in a pool to be reused to serve multiple connections. Default is `true`.
- `minNbOfThreads`: minimum number of threads to keep in the pool (if `poolThreads` is set to `true`). Default is 0.
- `maxNbOfThreads`: maximum number of threads in the pool (if `poolThreads` is set to `true`). Default is 0 (no limit).
- `maxThreadIdleTime`: maximum time in seconds a thread can remain idle before being removed from the pool. Default is 60 seconds (a thread that has not serve any request in the past 60 seconds will be killed).
- `sqlDumpLength`: maximum number of characters of a SQL statement to display in traces and exception messages. 0 means no limit and the full statement is inserted in the message (be careful especially if you are using large objects. Default is 40.
- `useStaticResultSetMetaData`: when fetching DatabaseMetaData, Sequoia returns built-in ResultSetMetaData that conforms to the JDBC standard (default set to true). If your application relies on specific extensions of the database driver, you can disable this feature so that ResultSetMetaData is always fetched from the driver (slower but transparent).

10.2.1. Distribution

A Distribution element defines the group communication settings when a virtual database is replicated by multiple controllers (feature called horizontal scalability). A Distribution element is defined as follows:

```
<!ELEMENT Distribution (MessageTimeouts)>
<!ATTLIST Distribution
  groupName              CDATA #IMPLIED
  hederaPropertiesFile   CDATA "/hedera_jgroups.properties"
>
```

- `MessageTimeouts`: Tunable message timeouts for messages sent to the group of controllers. The timeouts are usually properly tuned by default and don't need to be modified.

  `groupName`: group name to be used by the JGroups communication layer. If no name is provided, the virtual database name is used instead.

  Note: The JGroups stack configuration is defined in `config/total-token.xml`. Refer to the JGroups documentation if you want to alter the JGroups configuration.

  Note: When a controller fails, all backends attached to it are automatically disabled. A full recovery process is then needed for these nodes. To learn more about this issue, read the horizontal scalability design document (http://c-jdbc.objectweb.org/current/doc/C-JDBC_horizontal_scalability.pdf).

- `hederaPropertiesFile`: file name of the Hedera configuration file to use to setup the group communication for this virtual database. The default configuration use JGroups, but this value can be changed to use Appia instead.

10.2.1.1. Configuring Hedera group communication

Hedera is a group communication wrapper that provides various high-level facilities such as distributed remote procedure calls. Hedera default configuration for Sequoia relies on the JGroups group communication library. You must specify the JGroups factory and the name of the JGroups configuration file in the Hedera configuration file usually called `hedera_jgroups.properties`. Here is an example of such file:

```
hedera.factory=org.continuent.hedera.factory.JGroupsGroupCommunicationFactory
hedera.channel.jgroups.config=/total-token.xml
hedera.channel.jgroups.fragmentSize=32000
```

You can change the configuration of Hedera to use Appia, instead of JGroups. To change the configuration of Hedera to use Appia, change the `hederaPropertiesFile` parameter of the `Distribution` element to `/hedera_appia.properties`. The default Appia configuration uses the file `appia.xml`. Here is an example of a configuration that uses Appia:

```
<Distribution hederaPropertiesFile="/hedera_appia.properties">
  <MessageTimeouts/>
</Distribution>
```

Note: It is possible to have one Hedera configuration file per virtual database. Therefore group communication settings are per virtual database.

10.2.1.2. Controller replication requirements

When you replicate a virtual database in multiple controllers, there are some rules that you must follow:

10.2.2. Monitoring

> **Note:** Warning! Monitoring can possibly lead to a memory leak and should only be used on a short period of time There is also a JMX method on the VirtualDatabaseMBean to set this on and off while online:
>
> ```
> void setMonitoringToActive(boolean active) throws VirtualDatabaseException;
> ```

Monitoring provides a generic section for different monitoring modules. At the moment, only "SQLMonitoring" is provided.

```
<!ELEMENT Monitoring (SQLMonitoring*)>

<!ELEMENT SQLMonitoring (SQLMonitoringRule*)>
<!ATTLIST SQLMonitoring
    defaultMonitoring (on | off) "on"
>
```

SQL Monitoring provides statistics (count, error, cache hits, timing) for SQL queries. It is possible to define rules to turn monitoring on or off for specific query patterns.

`defaultMonitoring`: defines the default rule if a request should be monitored (on) or not (off) if no specific rule matches the request.

```
<!ELEMENT SQLMonitoringRule EMPTY>
<!ATTLIST SQLMonitoringRule
    queryPattern    CDATA #REQUIRED
    caseSensitive   (true | false) "false"
    applyToSkeleton (true | false) "false"
    monitoring      (on | off) "on"
>
```

A SQLMonitoringRule Defines a specific monitoring rule for all queries that match the given pattern.

- `queryPattern`: a regular expression understood by the Jakarta Regexp API. For more information on Regexp format, go to Jakarta Regexp web site (http://jakarta.apache.org/regexp).
- `caseSensitive`: true if the pattern matching must be case sensitive.
- `applytoSkeleton`: true if the pattern must apply to the query skeleton (found in PrepareStatement), false if the instantiated query should be used. Example: - *skeleton*: `SELECT * FROM t WHERE x=?` - *instantiated query*: `SELECT * FROM t WHERE x=12`
- `monitoring`: on to activate the monitoring for this rule, off to disable it.

Examples:

- `<SQLMonitoring queryPattern="^delete" monitorRequest="off"/>` will turn monitoring off for all delete queries.
- `<SQLMonitoring queryPattern="select * from users *" monitorRequest="on"/>` will turn monitoring on for all select queries on the users table.

> **Note:** *!Warning!* This is different from `<SQLMonitoring queryPattern="select \* from users *" monitorRequest="on"/>` which turns monitoring on for the "select * from users ..." kind of queries.

10.3. Backup Manager

A Backup Manager defines a number of Backuper in charge of performing backup/restore operations on backends. This element is defined as follows:

```
<!ELEMENT Backup (Backuper+)>

<!ELEMENT Backuper EMPTY>
<!ATTLIST Backuper
    backuperName CDATA #REQUIRED
    className    CDATA #REQUIRED
    options      CDATA #IMPLIED
>
```

A Backuper is defined by a logical backuperName used by the administration console when performing a backup operation. The className specifies the implementation of the Backuper. Backuper specific options can be provided as well (this can be the path to a properties files or a set of attributes). Check your Backuper documentation for its specific options.

Here is an example to use the Octopus Backuper for a virtual database:

```
<Backup>
  <Backuper backuperName="Octopus" className="org.continuent.sequoia.controller.backup.backupers.
</Backup>
```

> **Note:** Octopus does not have access to the Sequoia classloader for drivers and therefore it needs database drivers to be accessible from the controller classpath. A good solution is to unjar the drivers in the drivers/ directory of Sequoia.
>
> Octopus dumps are by default stored in a compressed .zip format.

10.4. Authentication Manager

An authentication manager element is defined as follows:

```
<!ELEMENT AuthenticationManager (Admin+, VirtualUsers)>

<!ELEMENT Admin (User+)>

<!ELEMENT User EMPTY>
<!ATTLIST User
    username CDATA #REQUIRED
    password CDATA #REQUIRED
>

<!ELEMENT VirtualUsers (VirtualLogin+)>
```

```
<!ELEMENT VirtualLogin (TrustedLogin*)>
<!ATTLIST VirtualLogin
    vLogin    CDATA #REQUIRED
    vPassword CDATA #REQUIRED
>
<!ELEMENT TrustedLogin EMPTY>
```

An authentication manager defines:

1. an administrator login to be used by the console to access the virtual database administration menu (see Section 8.5) that allows enabling the backends,
2. "virtual logins" that are used by the client application and that are mapped to "real logins" for each backend.
3. "trusted logins" will be used in the future to allow reusing other form of authentication from within Sequoia.

Here is an example of an authentication manager definition:

```
    <AuthenticationManager>
      <Admin>
       <User username="admin" password="adminPwd"/>
      </Admin>
      <VirtualUsers>
        <VirtualLogin vLogin="user1" vPassword="userPwd1"/>
        <VirtualLogin vLogin="user2" vPassword=""/>
      </VirtualUsers>
    </AuthenticationManager>
```

In this example, the virtual database has one administrator. The admin can use the login/password "admin/adminPwd" to log in the console.

Two virtual logins are defined: `user1` and `user2` with `userPwd1` and no password, respectively. These logins are those used by the client application and given to the Sequoia driver.

The connection manager to use with each of the virtual has to be defined in the DatabaseBackend section. Each DatabaseBackend has to define a pool connection manager for each of the virtual user specified here.

10.5. Database Backend

Each database backend must be given a unique name (it is a logical name but it is convenient to use the same name as the real machine name). The database schema is automatically gathered from the backend when it is added to the virtual database. However, you can specify a static database schema (refer to Section 10.5.2) to be used instead. Finally, a specific connection manager (see Section 10.5.3) defines the connection pooling strategy for each virtual login on each backend.

A database backend element is defined as follows:

```
<!ELEMENT DatabaseBackend (DatabaseSchema?, RewritingRule*, ConnectionManager+)>
<!ATTLIST DatabaseBackend
    name                     CDATA #REQUIRED
    driver                   CDATA #REQUIRED
    driverPath               CDATA #IMPLIED
    url                      CDATA #REQUIRED
    connectionTestStatement  CDATA #REQUIRED
    nbOfBackendWorkerThreads CDATA "5"
```

>

Here is a brief description of database backend attributes:

- `name`: the unique logical name identifying this backend.
- `driver`: the database native JDBC driver class name.
- `driverPath`: name of the directory or jar file containing the native driver files. If driverPath is omitted, the driver must be in the `drivers/` directory. If several driver jar files are in the same directory, the first jar file containing the class name specified in the `driver` attribute is used. Note that drivers are loaded in separate classloaders which allows you to use different versions of the same driver on different backends just by specifying the right jar file.
- `url`: the JDBC URL to connect to this database backend.
- `connectionTestStatement`: SQL statement to send on a connection to check if the connection is still valid. This is used when Sequoia suspects a connection to be broken after the failure of a request. This statement should not update the database because if the connection is still valid the database state should remain the same. Here are the settings for the most popular databases:
  - for MySQL use **select 1**
  - for PostgreSQL use **select now()**.
  - for Apache Derby use **values 1**.
  - for HSQL use **call now()**.
  - for SAP DB (MySQL MaxDB) use **select count(*) from versions**.
  - for Oracle use **select * from dual**.
  - for Firebird use **select 1 from rdb$types**.
  - for InstantDB use **set date format "yyyy/mm/dd"**.
  - for Interbase use **select * from rdb$types**.
  - for Microsoft SQL server 2000 **select 1**.

- `nbOfBackendWorkerThreads`: defines the number of BackendWorkerThread that can process writes in parallel (minimum is 2, default is 5). A large number of threads will generally not improve write performance.

Here is a complete example of a database backend element including its connection manager definition:

```
<DatabaseBackend name="node1" driver="org.gjt.mm.mysql.Driver"
      url="jdbc:mysql://node1.objectweb.org/rubis" connectionTestStatement="select 1">
  <ConnectionManager vLogin="user1" rLogin="ruser1" rPassword="rpass1">
    <SimpleConnectionManager/>
  </ConnectionManager>
  <ConnectionManager vLogin="user2">
    <VariablePoolConnectionManager initPoolSize="10"
                                   minPoolSize="5"
                                   maxPoolSize="100"/>
  </ConnectionManager>
</DatabaseBackend>
```

10.5.1. Rewriting requests on backends

If your cluster is made of database engines from different vendors, client requests might not be understood by all database backends. If your application was written for PostgreSQL and you want to add MySQL backends, some request might have to be adapted to execute correctly on MySQL. You can specify rules to rewrite queries on the fly on a specific backend. A RewritingRule element defines how a query matching a given pattern should be rewritten.

```
<!ELEMENT RewritingRule EMPTY>
<!ATTLIST RewritingRule
    queryPattern  CDATA #REQUIRED
    rewrite       CDATA #REQUIRED
    matchingType  (simple | pattern) "simple"
    caseSensitive (true | false) "false"
    stopOnMatch   (true | false) "false"
>
```

- `queryPattern`: SQL query pattern to match.
- `rewrite`: rewritten SQL query.
- `matchingType`: `simple`: means that the first occurrence of queryPattern in the request will be replaced by the string specified in rewrite. `pattern`: uses a pattern based match/replace. A pattern uses `?x` where x is a logical number assigned to the pattern. Example: `select ?1 from ?2 where x=?3`.
- `caseSensitive`: true if matching must be case sensitive.
- `stopOnMatch`: rules are applied in the order they are defined. If one rule matches and stopOnMatch is set to true, next rules are ignored. If stopOnMatch is set to false, if another rule matches the rewritten query, the query will be rewritten again.

Examples:

```
 <RewritingRule queryPattern="from user" rewrite="from "user;""
    matchingType="simple"/>
```

will rewrite the query `select * from user where x=y` as `select * from "user" where x=y`.

```
 <RewritingRule queryPattern="select * from t where x=?1"
    rewrite="select x from y where y=?1" matchingType="pattern"/>
```

will rewrite the query `select * from t where x=435` to `select x from y where y=435`

```
 <RewritingRule queryPattern="?1 LIMIT ?2,?3" rewrite="?1 LIMIT ?3,?2"
    matchingType="pattern"/>
```

will rewrite the query `select * from t limit 10,20` to `select * from t limit 20,10`

10.5.2. Database Schema Definition

DatabaseSchema groups static and dynamic definitions for gathering, constructing and validating the in-memory schema used for load balancing and caching.

A Database schema is defined as follow

```
<!ELEMENT DatabaseSchema (DatabaseStaticSchema?)>
```

```
<!ATTLIST DatabaseSchema
 dynamicPrecision (static|table|column|procedures|all) "all"
 gatherSystemTables (true | false) "false"
 schemaName CDATA #IMPLIED
>
```

- `dynamicSchemaPrecision`: if set to `static`, the controller will not check schemas and stored procedures, it will entirely rely on the statically defined schema. If set to something else than "static" it will get information from the backend to check validity of static schema at given level. `table` level will check for table names only, `column` level will check for column names, `procedures` will gather all executable stored procedures. All, includes all information that can be collected.
- `gatherSystemTables`: true if system tables and views should be retrieved, false otherwise (default).
- `schemaName`: if no schemaName is specified all objects visible to the user are gathered, otherwise only the objects belonging to the specified schema are used.

> **Note:** Default option for constructing the schema is to collect all information, even if a static schema is defined especially for checking validity of input. This can be really slow if the database has quite a number of stored procedures defined.

A static database schema can be defined to override the schema automatically gathered by the controller. However, the schema must remain compatible with the schema gathered from the backend.

A database schema element is defined as follows:

```
<!ELEMENT DatabaseStaticSchema (DatabaseProcedure*,DatabaseTable+)>

<!ELEMENT DatabaseProcedure (DatabaseProcedureParameter*)>
<!ATTLIST DatabaseProcedure
     name       CDATA #REQUIRED
     returnType (resultUnknown | noResult | returnsResult) "resultUnknown"
>

<!ELEMENT DatabaseTable (DatabaseColumn+)>
<!ATTLIST DatabaseTable
    tableName   CDATA #REQUIRED
    nbOfColumns CDATA #REQUIRED
>

<!ELEMENT DatabaseColumn EMPTY>
<!ATTLIST DatabaseColumn
    columnName CDATA #REQUIRED
    isUnique   (true | false) "false"
>
```

The `isUnique` attribute should be set to `true` if the column has a `UNIQUE` constraint. This is the case for primary keys (composed primary keys are not yet supported). This affects only cache behavior and select statements parsing.

Here is an example of a database schema definition:

```
<DatabaseStaticSchema>
```

```
   <DatabaseTable tableName="users" nbOfColumns="10">
      <DatabaseColumn columnName="id" isUnique="true"/>
      <DatabaseColumn columnName="firstname" isUnique="false"/>
      <DatabaseColumn columnName="lastname" isUnique="false"/>
      <DatabaseColumn columnName="nickname" isUnique="false"/>
      <DatabaseColumn columnName="password" isUnique="false"/>
      <DatabaseColumn columnName="email" isUnique="false"/>
      <DatabaseColumn columnName="rating" isUnique="false"/>
      <DatabaseColumn columnName="balance" isUnique="false"/>
      <DatabaseColumn columnName="creation_date" isUnique="false"/>
      <DatabaseColumn columnName="region" isUnique="false"/>
   </DatabaseTable>

   <DatabaseTable tableName="regions" nbOfColumns="2">
     <DatabaseColumn columnName="id" isUnique="true"/>
     <DatabaseColumn columnName="name" isUnique="false"/>
   </DatabaseTable>
</DatabaseStaticSchema>
```

10.5.3. Connection Manager

One connection manager must be defined for each virtual login (vLogin) the backend belongs to. The real user login/password (rLogin/rPassword) combination used to connect to the physical database backend is set by default to the same as the virtual login/password. An example of a connection manager definition is available in Section 10.5.

The connection manager element complete definition is as follows:

```
<!ELEMENT ConnectionManager (SimpleConnectionManager |
                             FailFastPoolConnectionManager |
                             RandomWaitPoolConnectionManager |
                             VariablePoolConnectionManager)>
<!ATTLIST ConnectionManager
    vLogin    CDATA #REQUIRED
    rLogin    CDATA #IMPLIED
    rPassword CDATA #IMPLIED
>

<!ELEMENT SimpleConnectionManager EMPTY>

<!ELEMENT FailFastPoolConnectionManager EMPTY>
<!ATTLIST FailFastPoolConnectionManager
    poolSize CDATA #REQUIRED
>

<!ELEMENT RandomWaitPoolConnectionManager EMPTY>
<!ATTLIST RandomWaitPoolConnectionManager
    poolSize CDATA #REQUIRED
    timeout  CDATA #IMPLIED
>

<!ELEMENT VariablePoolConnectionManager EMPTY>
<!ATTLIST VariablePoolConnectionManager
    initPoolSize CDATA #REQUIRED
    minPoolSize  CDATA #IMPLIED
```

```
    maxPoolSize  CDATA #IMPLIED
    idleTimeout  CDATA #IMPLIED
    waitTimeout  CDATA #IMPLIED
>
```

Sequoia offers several connection managers that are described hereafter:

- `SimpleConnectionManager`: basic connection manager that opens a new connection on each request and closes it at the end. It is useful if the underlying driver already implements connection pooling for example.
- `FailFastPoolConnectionManager`: offers connection pooling and fails fast when the pool is empty. `poolSize` is the size of the pool.

  All connections are initialized at startup time and if the pool size is too large it is adjusted to the largest number of connections available. Once the pool is empty, `null` is returned instead of a connection. Therefore incoming requests will fail until at least one connection is freed. No system overload should occur with this connection manager, but if the pool size is too small, many requests will fail.

- `RandomWaitPoolConnectionManager`: provides connection pooling and wait when the pool is empty until a connection is freed. This connection manager accepts the following attributes:
  - `poolSize`: this is the size of the pool.
  - `timeout`: this is the maximum time in seconds to wait for a connection to be freed. Default is 0 and means no timeout, that is to say that we wait until one connection is freed.

  All connections are initialized at startup time and if the pool size is too large it is adjusted to the largest number of connections available. Once the pool is empty, the requests wait until a connection is freed or the specified timeout has elapsed. The FIFO[3] order of connection request is not ensured by this connection manager since it relies on the Java wait/notify mechanism.

- `VariablePoolConnectionManager`: provides connection pooling with a dynamically adjustable pool size. This connection manager accepts the following attributes:
  - `initPoolSize`: initial pool size to be initialized at startup.
  - `minPoolSize`: minimum number of connections to keep in the pool. Default is equal to `initPoolSize`.
  - `maxPoolSize`: maximum number of connections in this pool. Default is 0 and means no limit.
  - `idleTimeout`: time in seconds a connection can stay idle before being released (removed from the pool). Default is 0 and means that once allocated, connections are never released.
  - `waitTimeout`: this is the maximum time in seconds to wait for a connection to be freed. Default is 0 and means no timeout, that is to say that we wait until one connection is freed.

10.6. Request Manager

The request manager is composed of a scheduler (see Section 10.6.2), an optional query cache (see Section 10.6.3), a load balancer (see Section 10.6.4) and an optional recovery log (see Section 10.6.5).

If requests need to be parsed, it can be done sequentially when needed (`backgroundParsing` is set to `false` which is the default value) or forced to be performed in background by a separate thread (it means a new thread is created for each request that need to be parsed).

Parsing is by default case insensitive (`caseSensitiveParsing` is set to `false`) which means that table and column names will be matched to the database schema without checking the case. If you want to enforce the parsing to be case sensitive and reject queries that do not use the same case for table and column names as the ones fetched from the database, set `caseSensitiveParsing` to `true`.

A timeout in seconds can be defined for begin/commit/rollback operations. If no value is given, the default timeout is set to 60 seconds. Warning! A value of 0 means no timeout and waits forever until completion.

The request manager element definition is as follows:

```
<!ELEMENT RequestManager (RequestScheduler, RequestCache?, LoadBalancer, RecoveryLog?)>
<!ATTLIST RequestManager
    backgroundParsing    (true | false) #IMPLIED
    caseSensitiveParsing (true | false) #IMPLIED
    beginTimeout      CDATA #IMPLIED
    commitTimeout     CDATA #IMPLIED
    rollbackTimeout   CDATA #IMPLIED
>
```

10.6.1. Macros Handler

Sequoia can interpret and replace on-the-fly macros with a value computed by the controller (the RequestManager in fact). This prevents different backends to generate different values when interpreting the macros which could result in data inconsistencies. The supported macros are the following:

- `rand`: RAND() can be replaced with an int, long, float or double value.

all the date macros (now, currentDate, currentTime, timeOfDay and currentTimestamp) can be replaced by one of the following:

- `off`: do not replace the macro

  `date`: java.sql.Date.toString() build from current time at controller (example: 2001-02-17)

  `time`: java.sql.Time.toString() build from current time at controller (example: 19:07:32).

  `timestamp`: java.sql.Timestamp.toString() build from current time at controller (example: 2001-02-17 19:07:32-05).

- `timeResolution`:defines the timer precision to use when rewriting a query that contains a date macro. Default is 0 millisecond which is the highest precision. A value of 1000 corresponds to a 1 second precision, 60000 to a 1 minute precision and so on.

The MacroHandling element definition is as follows:

```
<!ELEMENT MacroHandling EMPTY>
<!ATTLIST MacroHandling
        rand (off | int | long | float | double) "float"
        now  (off | date | time | timestamp) "timestamp"
        currentDate (off | date | time | timestamp) "date"
        currentTime (off | date | time | timestamp) "time"
        timeOfDay (off | date | time | timestamp) "timestamp"
```

```
        currentTimestamp (off | date | time | timestamp) "timestamp"
        timeResolution CDATA "0"
>
```

Note: A default Macrohandling element is instantiated and used if nothing is specified in the configuration file.

10.6.2. Request Scheduler

The request scheduler is responsible for scheduling the requests and ensuring a serializable execution order. Different schedulers are provided for each RAIDb level (see Section 9). Optimized schedulers are also provided for use with a single database backend (SingleDB configuration).

The request scheduler element definition is as follows:

```
<!ELEMENT RequestScheduler (SingleDBScheduler | RAIDb-0Scheduler |
                            RAIDb-1Scheduler | RAIDb-2Scheduler)>

<!ELEMENT SingleDBScheduler EMPTY>
<!ATTLIST SingleDBScheduler
    level (passThrough | pessimisticTransaction) #REQUIRED
>

<!ELEMENT RAIDb-0Scheduler EMPTY>
<!ATTLIST RAIDb-0Scheduler
    level (passThrough) #REQUIRED
>

<!ELEMENT RAIDb-1Scheduler EMPTY>
<!ATTLIST RAIDb-1Scheduler
    level (passThrough) #REQUIRED
>

<!ELEMENT RAIDb-2Scheduler EMPTY>
<!ATTLIST RAIDb-2Scheduler
    level (passThrough) #REQUIRED
>
```

Here is a brief definition of the meaning of each scheduler:

- `passThrough`: queries are just assigned a unique identifier and forwarded as is to the load balancer letting each database perform the scheduling and the locking. Therefore you will obtain the locking granularity provided by the database which should be row-level locking. The load balancer will just ensure that the writes are sent in the same order to all backends.
- `pessimisticTransaction`: this is a pessimistic transactional level scheduler that schedules transactions in a safe way (without possible deadlocks) but providing less parallelism for writes compared to optimistic scheduling (this is only sensitive on write heavy workloads).

10.6.3. Request Cache

A Request Cache can be composed of different caches that differ in the type of data they cache:

- `MetadataCache:` this cache improves the ResultSet creation time by keeping the various field information with their metadata. It is strongly encouraged to use this cache that reduces both CPU and memory usage.
- `ParsingCache:` allows to parse a request only once for all its executions. This reduces the CPU load on the controller.
- `ResultCache:` this cache keeps the results associated to a given request. Cache entries can be invalidated according to various policies. This cache reduces the load on the database backends.

A RequestCache element is defined as follows:

```
<!ELEMENT RequestCache (MetadataCache?, ParsingCache?, ResultCache?)>
```

10.6.3.1. Metadata Cache

The MetadataCache caches ResultSet metadata and fields meta information associated to a query execution so that each time a query is executed, we don't have to gather all metadata from the underlying driver and we can build Sequoia ResultSet much more efficiently both in terms of speed and memory usage. Note that if you use PreparedStatements, the query skeleton is used for matching the cache instead of the instantiated query.

Example: `SELECT * FROM t WHERE x=?` hits on the same cache entry for all queries of this form for any value of x.

A MetadataCache element is defined as follows:

```
<!ELEMENT MetadataCache EMPTY>
<!ATTLIST MetadataCache
    maxNbOfMetadata CDATA "10000"
    maxNbOfField    CDATA "0"
>
```

`maxNbOfMetadata:` maximum number of metadata entries to keep in the cache (default is 10000 and 0 means unlimited)

`maxNbOfField:` maximum number of field entries to keep in the cache (0 means unlimited and is the default setting).

10.6.3.2. Parsing Cache

Parsing requests is a resource consuming process. The ParsingCache caches the result of the parsing processing so that a request is only parsed once for all its executions. If you are using PreparedStatements, the ParsingCache can store the query skeleton meaning that the cached parsing will match any instances of the skeleton.

Example: `SELECT * FROM t WHERE x=?` will be parsed only once for any value of x.

A MetadataCache element is defined as follows:

```
<!ELEMENT ParsingCache EMPTY>
<!ATTLIST ParsingCache
    backgroundParsing (true | false) "false"
    maxNbOfEntries    CDATA "5000"
>
```

Request parsing can be done sequentially when needed (backgroundParsing set to false which is the default value) or forced to be performed in background by a separate thread (it means a new thread is created for each request that need to be parsed).

`maxNbOfEntries:` Defines the maximum number of entries to keep in the cache. The cache uses a LRU (Least Recently Used) replacing policy meaning that the oldest entries from the cache are removed when it is full. Default is 0 and means no limit on cache size.

10.6.3.3. Result Cache

The ResultCache caches results of queries. A query and its ResultSet are stored in the cache so that if the sames query is executed, the ResultSet stored in the cache is returned.

ResultCacheRule elements define the cache coherency and policy. Default cache behavior is eager consistency for all queries (ResultSet returned by the cache are always coherent and up-to-date). See below (ResultCacheRule element) to relax the cache consistency to achieve better performance.

> **Note:** Note that ResultSet caching is disabled if no result cache element is found in the configuration file

If two exact same requests are to be executed at the same time, only one is executed and the second one waits until the completion of the first one (this is the default `pendingTimeout` value which is 0). To prevent the second request from waiting forever, a `pendingTimeout` value in seconds can be defined for the waiting request. If the timeout expires, the request is executed in parallel with the first one.

A result cache element is described as follows:

```
<!ELEMENT ResultCache (DefaultResultCacheRule?, ResultCacheRule*)>
<!ATTLIST ResultCache
    granularity   (database | table | column | columnUnique) "database"
    maxNbOfEntries CDATA "100000"
    pendingTimeout CDATA "0"
>
```

The result cache `granularity` defines how entries are removed from the cache. `database` flush the whole cache on every write access. This is the default cache setting. `table` and `column` provide table-based and column-based invalidations, respectively. `columnUnique` can optimize requests that select a unique primary key (useful with EJB entity beans).

You can specify the maximum number of entries (default is 100000) to limit the cache size. This is obviously not as efficient as setting a cache size, but in the latter case we would have to spend a lot of time computing size of result sets from queries (Java does not provide a sizeof operator!). We offer size display in bytes when viewing the cache from the console though.

Finer grain tuning of the cache is based on rules matching query pattern. A queryPattern are regular expressions to match according to Jakarta Regexp (see their web site (http://jakarta.apache.org/regexp) for more information). A default cache rule defines the policy if no other rule matches:

```
<!ELEMENT DefaultResultCacheRule (NoCaching | EagerCaching | RelaxedCaching)>
<!ATTLIST DefaultResultCacheRule
    timestampResolution CDATA "1000"
>

<!ELEMENT ResultCacheRule (NoCaching | EagerCaching | RelaxedCaching)>
<!ATTLIST ResultCacheRule
```

```
    queryPattern        CDATA #REQUIRED
    caseSensitive       (true | false) "false"
    applyToSkeleton     (true | false) "false"
    timestampResolution CDATA "1000"
>
```

- `queryPattern`: the regular expression to match.
- `caseSensitive`: true if the pattern matching must be case sensitive
- `applytoSkeleton`: true if the pattern must apply to the query skeleton (found in PrepareStatement), false if the instantiated query should be used. Example: skeleton is `SELECT * FROM t WHERE x=?` and instantiated query is `SELECT * FROM t WHERE x=12`.
- `timestampResolution`: If a query contains a NOW() macro, it is replaced with the current date on the controller. timestampResolution indicates the resolution (in milliseconds) to use when replacing NOW() with the current date. If the resolution is below 1 second (value <1000ms), the request is never kept in the cache because there is almost no chance that the same request will come with the same timestamp. Note that this timestamp is for the cache only and you can use a greater resolution for the load balancer (see below).

  Note: If timestampResolution is set to 60000, every execution of a query like `SELECT * FROM x WHERE date=NOW()` will be replaced with the same value for 1 minute (i.e. `SELECT * FROM x WHERE date="2012-11-15 08:03:00.000"`) and therefore the cache entry may be hit for 1 minute.

To define a default rule that disable caching use:

```
<DefaultResultCacheRule>
  <NoCaching/>
</DefaultResultCacheRule>
```

If no default rule is provided, the following default rule is assumed:

```
<ResultCacheRule queryPattern="default" timestampResolution="1000">
  <EagerCaching/>
</ResultCacheRule>
```

Each cache rule can have a different caching behavior. The available behavior are the following:

```
<!ELEMENT NoCaching EMPTY>
<!ELEMENT EagerCaching EMPTY>
<!ELEMENT RelaxedCaching EMPTY>
<!ATTLIST RelaxedCaching
     timeout        CDATA "60"
     keepIfNotDirty (true | false) "true"
>
```

- `NoCaching` means we do not put the match in the cache
- `EagerCaching` means that all entries in the cache are always coherent and any update query (insert,delete,update,...) will automatically invalidate the corresponding entry in the cache. This was the previous cache behavior for all queries

- `RelaxedCaching` means that a `timeout` is set for this entry and the entry is kept in the cache until the timeout expires. When the timeout expires, if no write has modified the corresponding result and `keepIfNotDirty` is set to true, the entry is kept in the cache and the timeout is rearmed (reset) with its initial value.

  Note: RelaxedCaching may provide stale data. The timeout defines the maximum staleness of a cache entry. It means that the cache may return an entry that is out of date.

Here is a cache rule example:

```
<ResultCacheRule queryPattern="select ? from b where id=?" applyToSkeleton="true">
  <RelaxedCaching timeOut="6000" keepIfNotDirty="true"/>
</ResultCacheRule>
```

10.6.4. Load Balancer

The load balancer defines the way requests will be distributed among the backends according to a RAIDb level (see Section 9). It is possible to enforce a specific transaction isolation level on all connections (note that this will have no effect if the underlying database does not support this transaction isolation). By default, the default transaction isolation level will be used and no specific isolation will be enforced on the connections. The following load balancers are available:

- `SingleDB`: load balancer for a single database backend instance. This is only available if you use a single controller.
- `ParallelDB`: load balancer to use with a parallel database such as Oracle Parallel Server or Middle-R. Both read and write are load balanced on the backends, letting the parallel database replicating writes.
- `RAIDb-0`: full database partitioning (no table can be replicated) with an optional policy specifying where new tables are created.
- `RAIDb-1`: full database mirroring (all tables are replicated everywhere) with an optional policy specifying how distributed queries (writes/commit/rollback) completion is handled (when the first, a majority or all backends complete).
- `RAIDb-1ec`: full database mirroring (like RAIDb-1) with error checking for byzantine failure detection.
- `RAIDb-2`: partial replication (each table must be at least replicated once) with optional policies for new table creation (like RAIDb-0) and distributed queries completion (like RAIDb-1).
- `RAIDb-2ec`: partial replication (like RAIDb-2) with error checking for byzantine failure detection.

The load balancer element definition is as follows:

```
<!ELEMENT LoadBalancer (SingleDB | ParallelDB | RAIDb-0 | RAIDb-1 | RAIDb-1ec | RAIDb-2 | RAIDb-2
<!ATTLIST LoadBalancer
   transactionIsolation (databaseDefault | readUncommitted | readCommitted | repeatableRead | seri
>
```

10.6.4.1. `SingleDB` load balancer

The `SingleDB` load balancer does not need any specific parameter. The definition of the `SingleDB` element is as follows:

```
<!ELEMENT SingleDB EMPTY>
```

10.6.4.2. `ParallelDB` load balancer

The `ParallelDB` load balancer must be used with a SingleDB request scheduler. This load balancer provides two implementations: `ParallelDB-RoundRobin` and `ParallelDB-LeastPendingRequestsFirst` providing round robin and least pending request first load balancing policies, respectively. ParallelDB load balancers are designed to provide load balancing and failover on top of parallel databases such as Oracle Parallel Server or Middle-R. It means that read and write requests are just sent to one alive backends, the parallel database being responsible for maintaining the consistency between the backends. The definition of the `ParallelDB` element is as follows:

```
<!ELEMENT ParallelDB (ParallelDB-RoundRobin | ParallelDB-LeastPendingRequestsFirst)>

<!ELEMENT ParallelDB-RoundRobin EMPTY>
<!ELEMENT ParallelDB-LeastPendingRequestsFirst EMPTY>
```

No specific settings are required for these load balancers. They do not require request parsing which means that requests are just forwarded as is to the backends (rewriting rules are still applied but no automatic transformation is performed).

10.6.4.3. `RAIDb-0` load balancer

The `RAIDb-0` load balancer accepts a policy to specify where new tables are created. The definition of the `RAIDb-0` element is as follows:

```
<!ELEMENT RAIDb-0 (MacroHandling?,CreateTable*)>

<!ELEMENT CreateTable (BackendName*)>
<!ATTLIST CreateTable
    tableName     CDATA #IMPLIED
    policy        (random | roundRobin | all) #REQUIRED
    numberOfNodes CDATA #REQUIRED
>

<!-- BackendName simply identifies a backend by its logical name -->
<!ELEMENT BackendName EMPTY>
<!ATTLIST BackendName
    name CDATA #REQUIRED
>
```

If `MacroHandling` is omitted, a default MacroHandling element is added.

`CreateTable` defines the policy to adopt when creating a new table. This policy is based on the given list of `BackendName` nodes (which might be a subset of the complete set of backends). If the backend list is omitted, then all enabled backends are taken at decision time. The attributes have the following meaning:

- `numberOfNodes` represents the number of backends to pickup from the `BackendName` list to apply the policy (it must be set to 1 for `RAIDb-0` load balancers and can never be greater than the number of nodes declared in the `BackendName` list).
- `policy` works as follows:

- `random`: `numberOfNodes` backends are picked up randomly from the `BackendName` list and the table is created on these nodes.
- `roundRobin`: `numberOfNodes` backends are picked up from the `BackendName` list using a round-robin algorithm and the table is created on these nodes.
- `all`: the table is created on *all* nodes in the `BackendName` list (`numberOfNodes` is ignored).

Here is an example of a `RAIDb-0` controller with three nodes where new tables are created randomly on the first two nodes:

```
...
<DatabaseBackend name="node1" ...
<DatabaseBackend name="node2" ...
<DatabaseBackend name="node3" ...
...

<LoadBalancer>
  <RAIDb-0>
    <CreateTable policy="random" numberOfNodes="1">
      <BackendName name="node1" />
      <BackendName name="node2" />
    </CreateTable>
  </RAIDb-0>
</LoadBalancer>
```

10.6.4.4. `RAIDb-1:full mirroring` load balancer

A RAIDb-1 load balancer is defined as follows:

```
<!ELEMENT RAIDb-1 (WaitForCompletion?, MacroHandling?, (RAIDb-1-RoundRobin |
           RAIDb-1-WeightedRoundRobin | RAIDb-1-LeastPendingRequestsFirst))>

<!ELEMENT RAIDb-1-RoundRobin EMPTY>
<!ELEMENT RAIDb-1-WeightedRoundRobin (BackendWeight)>
<!ELEMENT RAIDb-1-LeastPendingRequestsFirst EMPTY>

<!ELEMENT WaitForCompletion EMPTY>
<!ATTLIST WaitForCompletion
    policy (first | majority | all) "first"
    deadlockTimeoutInMs CDATA "30000"
>

<!ELEMENT BackendWeight EMPTY>
<!ATTLIST BackendWeight
    name   CDATA #REQUIRED
    weight CDATA #REQUIRED
>
```

If `WaitForCompletion` is omitted, the default behaviour is to return the result as soon as one backend has completed. deadlockTimeout defines the time in milliseconds to wait before starting the deadlock detection computation. This should typically be larger than the database deadlock timeout settings. Default is 30000 (30 seconds) but should typically be larger than the largest query execution time. 0 disables the deadlock detection.

If MacroHandling is omitted, a default MacroHandling element is added.

The `RAIDb-1` load balancer accepts a policy to specify distributed queries completion. Several load balancing policies are proposed:

- `RoundRobin`: simple round-robin load balancing. The first request is sent to the first node, the second request to the second node, etc... Once a request has been sent to the last backend, the next request is sent to the first backend and so on.
- `WeightedRoundRobin`: same as round-robin but a weight is associated to each backend. A backend that has a weight of 2 will get two times more requests than a backend with a backend with a weight of 1.
- `LeastPendingRequestsFirst`: the request is sent to the backend that has the least pending requests to execute (that can be considered as the shortest pending request queue).

The definition of the `RAIDb-1` element is as follows:

`WaitForCompletion` defines the policy to adopt when waiting for the completion of a request. Policy works as follows:

- `first`: returns the result as soon as one node has completed.
- `majority`: returns the result as soon as a majority of nodes (n/2+1) has completed.
- `all`: waits for all nodes to complete before returning the result to the client.

10.6.4.5. `RAIDb-1ec` load balancer

The RAIDb-1 with error checking must provide an error checking policy (defined below). The optional `WaitForCompletion` policy only concern write requests (`INSERT`, `DELETE`, `UPDATE`, commit, ...).

> **Note:** `RAIDb-1ec` is not operational in Sequoia v1.0alpha.

The definition of the `RAIDb-1ec` element is as follows:

```
<!ELEMENT RAIDb-1ec (WaitForCompletion?, ErrorChecking, (RAIDb-1ec-RoundRobin |
                     RAIDb-1ec-WeightedRoundRobin))>
<!ATTLIST RAIDb-1ec
    nbOfConcurrentReads CDATA #REQUIRED
>

<!ELEMENT RAIDb-1ec-RoundRobin EMPTY>
<!ELEMENT RAIDb-1ec-WeightedRoundRobin (BackendWeight)>

<!ELEMENT ErrorChecking EMPTY>
<!ATTLIST ErrorChecking
    policy        (random | roundRobin | all) #REQUIRED
    numberOfNodes CDATA #REQUIRED
>
```

Error checking policy (for RAIDb-1ec and RAIDb2-ec). Error checking is used to detect byzantine failures of nodes. It means detecting when a node sends funny results in a non-deterministic way. Error checking allows read queries to be sent to more than one database, and the results are compared. A majority of nodes must agree on the result that will be sent to the client. Error checking policies are defined as follows:

- `random`: `numberOfNodes` backends are picked up randomly; the read request is sent to these backends and results are compared.
- `roundRobin`: `numberOfNodes` backends are picked up using a round-robin algorithm ; the read request is send to these backends and results are compared.
- `all`: the request is sent to *all* nodes (`numberOfNodes` is ignored) and the results compared.

`numberOfNodes` must be greater or equal to 3.

10.6.4.6. `RAIDb-2 : distributed mirroring` load balancer

The definition of the `RAIDb-2` element is as follows:

```
<!ELEMENT RAIDb-2 (CreateTable*, WaitForCompletion?, MacroHandling?, (RAIDb-2-RoundRobin |
 RAIDb-2-WeightedRoundRobin | RAIDb-2-LeastPendingRequestsFirst))>


<!ELEMENT RAIDb-2-RoundRobin EMPTY>
<!ELEMENT RAIDb-2-WeightedRoundRobin (BackendWeight)>
<!ELEMENT RAIDb-2-LeastPendingRequestsFirst EMPTY>
```

If MacroHandling is omitted, a default MacroHandling element is added.

The `RAIDb-2` load balancer accepts a policy to specify where new tables are created and how distributed queries completion should be handled. Several load balancing policies are proposed:

- `RoundRobin`: simple round-robin load balancing. The first request is sent to the first node, the second request to the second node, etc... Once a request has been sent to the last backend, the next request is sent to the first backend and so on.
- `WeightedRoundRobin`: same as round-robin but a weight is associated to each backend. A backend that has a weight of 2 will get two times more requests than a backend with a backend with a weight of 1.
- `LeastPendingRequestsFirst`: the request is sent to the backend that has the least pending requests to execute (that can be considered as the shortest pending request queue).

The `CreateTable` element definition is defined in Section 10.6.4.3.

The `WaitForCompletion` element definition is defined in Section 10.6.4.4.

10.6.4.7. `RAIDb-2ec` load balancer

The RAIDb-2 with error checking must provide an error checking policy as in `RAIDb-1ec` (see Section 10.6.4.5). The other elements are similar to those defined for `RAIDb-2` controller (see Section 10.6.4.6).

> **Note:** `RAIDb-2ec` is not operational in Sequoia v1.0alpha.

The definition of the `RAIDb-2ec` element is as follows:

```
<!ELEMENT RAIDb-2ec (CreateTable*, WaitForCompletion?, ErrorChecking,
                     (RAIDb-2ec-RoundRobin | RAIDb-2ec-WeightedRoundRobin))>
<!ATTLIST RAIDb-2ec
    nbOfConcurrentReads CDATA #REQUIRED
>
```

```
<!ELEMENT RAIDb-2ec-RoundRobin EMPTY>
<!ELEMENT RAIDb-2ec-WeightedRoundRobin (BackendWeight)>
```

10.6.5. Recovery Log

The Sequoia Recovery Log stores write queries and transactions between logical checkpoints defined by the user. The log can be only be stored in a database (or cluster of databases) using a `JDBCRecoveryLog` element.

The definition of a `RecoveryLog` element is as follows:

```
<!ELEMENT RecoveryLog (JDBCRecoveryLog)>
```

10.6.5.1. Recoverylog

The RecoveryLog stores the recovery information in a database. To access this database, you must provide the driver class name to load (driver), an optional jar file or directory where to find the class to load (driverPath), the JDBC url to access the database as well as a valid login/password.

A timeout in seconds can be defined for the sql requests. If no value is given, the default timeout is set to 60 seconds. Warning! 0 means no timeout and wait forever until completion.

recoveryBatchSize is used to speedup the recovery process and allow several queries to be accumulated into a batch on the recovering backend. Increasing this value beyond a certain limit will not increase performance and will consume a significant amount of memory. Default is 10 and minimum is 1.

The recovery information is stored in 4 tables defined in the RecoveryLogTable, CheckpointTable, BackendLogTable and DumpTable elements.

The definition of a `RecoveryLog` element is as follows:

```
<!ELEMENT RecoveryLog (RecoveryLogTable, CheckpointTable, BackendTable, DumpTable)>

<!ATTLIST RecoveryLog
    driver            CDATA #REQUIRED
    driverPath        CDATA #IMPLIED
    url               CDATA #REQUIRED
    login             CDATA #REQUIRED
    password          CDATA #REQUIRED
    requestTimeout    CDATA "60"
    recoveryBatchSize CDATA "10"
>
<!ELEMENT RecoveryLogTable EMPTY>
<!ATTLIST RecoveryLogTable
    createTable              CDATA "CREATE TABLE"
    tableName                CDATA "logtable"
    logIdColumnType          CDATA "BIGINT NOT NULL"
    vloginColumnType         CDATA "VARCHAR NOT NULL"
    sqlColumnName            CDATA "sql"
    sqlColumnType            CDATA "VARCHAR NOT NULL"
    autoConnTranColumnType   CDATA "CHAR(1) NOT NULL"
    transactionIdColumnType  CDATA "BIGINT NOT NULL"
    requestIdColumnType      CDATA "BIGINT"
    execTimeColumnType       CDATA "BIGINT"
    updateCountColumnType    CDATA "INT"
    extraStatementDefinition CDATA ",PRIMARY KEY (log_id)"
```

```
>

<!ELEMENT CheckpointTable EMPTY>
<!ATTLIST CheckpointTable
    createTable              CDATA "CREATE TABLE"
    tableName                CDATA "checkpointtable"
    checkpointNameColumnType CDATA "VARCHAR NOT NULL"
    logIdColumnType          CDATA "BIGINT"
    extraStatementDefinition CDATA ",PRIMARY KEY (name)"
>

<!ELEMENT BackendTable EMPTY>
<!ATTLIST BackendTable
    createTable                 CDATA "CREATE TABLE"
    tableName                   CDATA "backendtable"
    databaseNameColumnType      CDATA "VARCHAR NOT NULL"
    backendNameColumnType       CDATA "VARCHAR NOT NULL"
    backendStateColumnType      CDATA "INTEGER"
    checkpointNameColumnType    CDATA "VARCHAR NOT NULL"
    extraStatementDefinition    CDATA ""
>

<!ELEMENT DumpTable EMPTY>
<!ATTLIST DumpTable
    createTable              CDATA "CREATE TABLE"
    tableName                CDATA "dumptable"
    dumpNameColumnType       CDATA "VARCHAR NOT NULL"
    dumpDateColumnType       CDATA "TIMESTAMP"
    dumpPathColumnType       CDATA "VARCHAR NOT NULL"
    dumpFormatColumnType     CDATA "VARCHAR NOT NULL"
    checkpointNameColumnType CDATA "VARCHAR NOT NULL"
    backendNameColumnType    CDATA "VARCHAR NOT NULL"
    tablesColumnName         CDATA "tables"
    tablesColumnType         CDATA "VARCHAR NOT NULL"
>
```

The `RecoveryLog` element requires the following attributes:

- `driver`: the driver class name
- `url`: the JDBC URL to access the database
- `login`: the user login to connect to the database
- `password`: the user password to connect to the database
- `requestTimeout`: optional timeout request in second that will be used to replay the log queries. Default timeout is 60 seconds and 0 means no timeout (wait forever until a request complete).
- `recoveryBatchSize`: used to speedup the recovery process and allow several queries to be accumulated into a batch on the recovering backend. Increasing this value beyond a certain limit will not increase performance and will consume a significant amount of memory. Default is 10 and minimum is 1.

The `RecoveryLogTable` defines how the JDBCRecoveryLog log table is created. The log table name (`tableName`) must conform to the syntax of a database table name. The log table stores a unique request id (`id`), the virtual login (`vlogin`)to use to execute the sql statement (`sql`) in the given transaction (`transactionId`). The statement used by the RecoveryLog to create the log table uses the RecoveryLogTable attributes as follows:

```
createTable tableName (
  log_id         logIdColumnType,
  vlogin         vloginColumnType,
  sqlColumnName  sqlColumnType,
  auto_conn_tran autoConnTranColumnType,
  transaction_id transactionIdColumnType,
  request_id     requestIdColumnType,
  exec_time      execTimeColumnType,
  update_count   updateCountColumnType,
  extraStatementDefinition)
```

If all default values are used, the log table is created using the following statement:

```
CREATE TABLE logtable (
  log_id         BIGINT NOT NULL UNIQUE,
  vlogin         VARCHAR NOT NULL,
  sql            VARCHAR NOT NULL,
  auto_conn_tran CHAR(1) NOT NULL,
  transaction_id BIGINT NOT NULL
  request_id     BIGINT,
  exec_time      BIGINT,
  update_count   INT,
  PRIMARY KEY (log_id)
  )
```

The `CheckpointTable` stores the checkpoint name and the corresponding index in the recovery log table. The statement used by the JDBCRecoveryLog to create the checkpoint table uses the CheckpointTable attributes as follows:

```
CREATE TABLE tableName (
  name   checkpointNameColumnType,
  log_id logIdColumnType
  extraStatementDefinition)
```

If all default values are used, the log table is created using the following statement:

```
CREATE TABLE checkpointtable (
  name   VARCHAR NOT NULL,
  log_id BIGINT,
  PRIMARY KEY(name))
```

The `BackendLogTable` stores the states of the different backends of a virtual database. It stores the name of the backend, the database it belongs to and the last known checkpoint of a backend when the backend is disabled, and the state the backend was in when the database was last shutdown. If all default values are used, the log table is created using the following statement:

```
CREATE TABLE backendtable (
  database_name   VARCHAR NOT NULL,
  backend_name    VARCHAR NOT NULL,
  backend_state   INTEGER,
  checkpoint_name VARCHAR NOT NULL
  )
```

Here is an example on how to define a JDBCRecoveryLog to work with a HSQL database:

```
<RecoveryLog>
   <RecoveryLog driver="org.hsqldb.jdbcDriver" url="jdbc:hsqldb:hsql://localhost" login="sa" passw
      <RecoveryLogTable
         tableName="recovery"
         logIdColumnType="INTEGER NOT NULL"
         sqlColumnType="VARCHAR NOT NULL"
        extraStatementDefinition=",PRIMARY KEY (id)"/>
      <CheckpointTable tableName="checkpoint"/>
  <BackendLogTable tableName="backendTable"/>
   </JDBCRecoveryLog>
</RecoveryLog>
```

The `DumpTable` stores the dump names and associated meta-data such as the corresponding checkpoint name. The statement used by the JDBCRecoveryLog to create the dump table uses the DumpTable attributes as follows:

```
  createTable tableName (
    dump_name       dumpNameColumnType,
    dump_date       dumpDateColumnType,
    dump_path       dumpPathColumnType,
    dump_format     dumpTypeColumnType,
    checkpoint_name checkpointNameColumnType,
    backend_name    backendNameColumnType,
    tables          tablesColumnType
    extraStatementDefinition)
```

dump_name is the dump logical name, dump_date the date at which the backup was started, dump_path the path where the dump can be found, dump_format an implementation specific text form that specifies the method used for the dump, checkpoint_name is the name of the checkpoint associated to this dump, tables is the list of tables that are contained in this dump (* means all tables). If all default values are used, the log table is created using the following statement:

```
  CREATE TABLE DumpTable (
    dump_name       VARCHAR NOT NULL,
    dump_date       TIMESTAMP,
    dump_path       VARCHAR NOT NULL,
    dump_format     VARCHAR NOT NULL,
    checkpoint_name VARCHAR NOT NULL,
    backend_name    VARCHAR NOT NULL,
    tables          VARCHAR NOT NULL
    )
```

10.7. SSL Configuration

SSL may be used for encryption as well as authentication for all connections to sequoia.

SSL support for sequoia is based on the Java Secure Socket Extension (JSSE). JSSE has been integrated into the Java 2 SDK, Standard Edition, v 1.4. For java 1.3 it can be installed as an optional package. (available at http://java.sun.com/products/jsse/)

10.7.1. Controller

On the controller side ssl can be configured for all jmx connections and the virtual database accessed via the sequoia driver with the xml element SSL in the controller configuration :

```
<!ELEMENT SSL EMPTY>
<!ATTLIST SSL
    keyStore              CDATA        #IMPLIED
    keyStorePassword      CDATA        #IMPLIED
    keyStoreKeyPassword   CDATA        #IMPLIED
    isClientAuthNeeded    (true|false) "false"
    trustStore            CDATA        #IMPLIED
    trustStorePassword    CDATA        #IMPLIED
>
```

- keyStore: The file where the keys are stored
- keyStorePassword: the password to the keyStore
- keyStoreKeyPassword: the password to the private key, if none is specified the same password as for the store is used
- isClientAuthNeeded: if set to false ssl is used for encryption only, it true set to true then the server is only accepting trusted clients (the client certificate has to be in the trusted store)
- trustStore: the file where the trusted certificates are stored, if none is specified the same store as for the key is used
- trustStorePassword: the password to the trustStore, if none is specified the same password as for the keyStore is used

10.7.2. Console / Jmx Clients

The console and other jmx clients are configured with the use of java properties :

- javax.net.ssl.keyStore
- javax.net.ssl.keyStorePassword
- javax.net.ssl.trustStore
- javax.net.ssl.trustStorePassword

Example : -Djavax.net.ssl.trustStore=client.keystore -Djavax.net.ssl.trustStorePassword=clientpassword

10.7.3. Driver

SSL on the driver side is configured with java properties

- sequoia.ssl.enabled=true
- javax.net.ssl.keyStore
- javax.net.ssl.keyStorePassword
- javax.net.ssl.trustStore
- javax.net.ssl.trustStorePassword

Example : -Djavax.net.ssl.keyStore=client.keystore -Djavax.net.ssl.keyStorePassword=clientpassword

10.7.4. Certificates (public and private keys)

You may create your certificates with the keytool (part of JSSE)

1. Create a self-signed server and a self-signed client key each in its own keystore

```
$> keytool -genkey -v -keyalg RSA -keystore server.keystore -dname "CN=Server, OU=Bar, O=Foo,
$> keytool -genkey -v -keyalg RSA -keystore client.keystore -dname "CN=Client, OU=Bar, O=Foo,
```

2. Export the server's and the client's public keys from their respective keystores

```
$> keytool -export -rfc -keystore server.keystore -alias mykey -file server.public-key
$> keytool -export -rfc -keystore client.keystore -alias mykey -file client.public-key
```

3. Import the client's public key to the server's keystore, and vice-versa:

```
$> keytool -import -alias client -keystore server.keystore -file client.public-key
$> keytool -import -alias server -keystore client.keystore -file server.public-key
```

10.8. Configuration Examples

Configuration files examples are available in the Sequoia distribution in the `/sequoia/doc/examples` directory.

Here is a brief overview of each example content:

- `Cache`: gives various configuration examples on how to use the cache.
- `Derby` : contains examples for the Apache Derby database including the ones used in the ApacheCon demos.
- `HorizontalScalbility` : provides configuration files to create a distributed virtual database on 2 controllers. One file should be loaded on each of the two controllers.
- `LinuxService` and `SuSE` : contains examples to run Sequoia controller as a Linux service.
- `SingleDB`: a Sequoia configuration with a unique MySQL backend.
- `RAIDb-0`: Sequoia configuration examples for RAIDb-0.
  - `RAIDb-0.xml`: a simple 2 nodes RAIDb-0 configuration.
  - `RAIDb-0-schema.xml`: a 2 nodes RAIDb-0 configuration using a static database schema definition matching the RUBiS benchmark database schema.
- `RAIDb-1`: contains various RAIDb-1 configuration examples.
- `RAIDb-2`: contains various RAIDb-2 configuration examples.
- `RecoveryLog`: gives an example of a fault tolerant recovery log.

Table 1. List of acronyms used in this document

C-JDBC	Clustered Java DataBase Connectivity
CVS	Concurrent Versions System
INRIA	French National Institute for Research in Computer Science and Control
JDBC	Java DataBase Connectivity (not officially recognized as such)
JMX	Java Management eXtensions
JRE	Java Runtime Environment
JVM	Java Virtual Machine
LGPL	GNU Lesser General Public License
RAIDb	Redundant Array of Inexpensive Databases
RDBMS	Relational DataBase Management System
RMI	Remote Method Invocation
SQL	Standard Query Language

11. Glossary

Table 1 summarizes all the acronyms used in this document.

12. About Sequoia

12.1. License

Sequoia is free software. You can redistribute it and/or modify it under the terms of the Apache v2 license (http://www.apache.org/licenses/LICENSE-2.0.html).

Sequoia is copyrighted by the French National Institute For Research In Computer Science And Control (http://www.inria.fr/) (INRIA) and Continuent.

12.2. Web Site

The Sequoia project is hosted on the Continuent.org web site at the following URL: http://sequoia.continuent.org/. To facilitate the development, a Sequoia project has also be created on the Continuent Forge (https://forge.continuent.org/). This GForge integrates with JIRA (https://forge.continuent.org/jira/browse/SEQUOIA) for bug tracking.

12.3. Mailing Lists

Two mailing lists are currently available for Sequoia. Both lists are archived for public review at the Sequoia's Web site (http://sequoia.continuent.org/).

- <sequoia@continuent.org> is the user mailing list. It is the source to get the latest information about Sequoia, send your feedback and get support from the Sequoia community.

- <sequoia-commits@continuent.org> is a developer mailing list that reports every commit in the Sequoia CVS repository.

Feedback is crucial to improve Sequoia. Please send us your comments or any other form of input to: <sequoia@continuent.org>.

12.4. Reporting a Bug

JIRA (https://forge.continuent.org/jira/browse/SEQUOIA) provides support for bug tracking. We strongly encourage you to use the automatic Report feature (see Section 7.3.3) that provides all the details we usually need to figure out what happened. If you cannot use this feature, please include the following information when reporting a bug (when applicable):

- The Sequoia driver and controller version.
- The XML file you used to configure the Sequoia controller.
- JDK vendor and version (example: Sun JDK 1.4.2_08). If you use different JDK for driver and controller, please give as much detail as possible.
- OS vendor and version (examples: Linux 2.6.12 or Windows XP® SP2). If you use different operating systems for clients, controllers and backends, give the appropriate information.
- Database backend version and driver (example: PostgreSQL 8.0.3 Linux with JDBC driver postgresql-8.0-312.jdbc3.jar).
- Detailed error description with possibly the exception stack trace or a logging trace with debugging enabled.

12.5. Getting Involved

Sequoia is an open source project and welcomes external contributions. Please read the Sequoia Developer's Guide and join us!

Basically, any feature you need but you do not find implemented in Sequoia may become a contribution topic. Simply send your ideas, documents and developments (if any) to the <sequoia@continuent.org> mailing list. Available tasks and the roadmap is available on JIRA (https://forge.continuent.org/jira/browse/SEQUOIA). Please use also this tool for feature requests and bug reports/fixes.

You can finally subscribe to the sequoia-commits mailing list if you want to receive notifications of the CVS changes.

12.6. About Continuent.org

Continuent.org (http://www.continuent.org/) Continuent.org is an open source portal and community dedicated to high availability and scalability services for databases and other closely related technologies. Continuent.org is sponsored by Continuent (http://www.continuent.com/).

12.7. About INRIA

INRIA (http://www.inria.fr/) is the French National Institute for Research in Computer Science and Control. The Sardes project (http://sardes.inrialpes.fr/) at INRIA Rhones-Alpes has defined the RAIDb concept and developed C-JDBC. Sequoia is a continuation of the C-JDBC project.

12.8. About ObjectWeb

The goal of the ObjectWeb Consortium (http://www.objectweb.org/) is the development of open source distributed middleware, in the form of adaptable and flexible components. ObjectWeb components range from specific software frameworks and protocols to integrated platforms. More information on ObjectWeb and its projects is available at the ObjectWeb's Web site.

Notes

1. Sequoia may work with older JVM version, but hasn't been tested.
2. CVS stands for *Concurrent Versions System* and is a popular version control system.
3. First In First Out.