Choose an Installation Scenario

This page describes the architectural choices to be made before configuring a NetarchiveSuite installation.

Contents

- 1 Choose a platform
- 2 Choose Repository
- 3 Choose the type of database
  - 3.1 Derby Database
  - 3.2 PostgreSQL Database
- 4 Choose a JMS broker
- 5 Java
- 6 Choose the set of machines taking part in the installation/deployment

Choose a platform

NetarchiveSuite can be installed in a number of different ways, with varying numbers of machines on different sites. There are a number of separate applications in play, most of which can be put on separate machines as needed. To keep clear what is necessary for which setups, we will consider the following types of setup:

- A. Single-machine setup. This corresponds to the setup used in the Quick Start Manual, where all applications run on the same machine, and file transfer are done by simply copying files locally. It is the simplest setup, but does not scale very well.
- B. Single-site setup. In this scenario, multiple machines are involved, necessitating file transfer between machines and multiple installations of the code. However, the machines are expected to be within the same firewall, so port setup should be no problem.
- C. Single-site setup with duplicate archive. This expands on the single-site set-up in that more than one copy of the archived files is used, using the concept of separate “Replica” to indicate the duplicates.
- D. Multi-site setup. When more than one site (physical location) is involved, separated by firewalls, extra issues of opening ports and specifying the correct site come into play. This is the most complex scenario, but also more secure against systematic errors, hacking, and other disasters.

Choose Repository

Scenario A and B from section Choose a platform involve having a local arcrepository without means of bitarchive replicas. This is configured by a plug-in (please refer to Configure Plugins in the Configuration Manual). In this scenario you would typically then use your own institutional bitpreservation solution to guarantee the long-term stability of the archived data.

Scenarios C and D from section Choose a platform involve having distributed bitarchive replicas. In these scenarios we have at least two bitarchive replicas. The Replica information must be configured before deployment either in the local settings file or included in the deploy configuration file for your system (please refer to Configure Repository in the Configuration Manual). (Note: the Danish Netarchive has a medium-term goal of migrating its backend bitpreservation to the Bitrepository system, so the distributed ArcRepository software currently included in NetarchiveSuite is liable to become obsolescent.)

Choose the type of database

The NetarchiveSuite can use two types of database:

- Derby database (mostly for testing)
- PostgreSQL (recommended for production)

However by default, NetarchiveSuite uses an external Derby server.

In fact NetarchiveSuite uses up to three distinct databases depending on which modules of NetarchiveSuite are deployed - a harvest database, an arcrepository administration database, and a wayback indexing database.

The wayback database is configured through hibernate and should be neutral as to which type of database is used. The harvest and admin databases are configured via the NetarchiveSuites settings file or by the use of deployment settings if the NetarchiveSuite deploy-application is used.

Derby is used in the QUICKSTART installation, but PostgreSQL is recommended for large installations where it should have superior performance and better support from external tools.

Derby Database

If you want to use a Derby database server, you have to start it as a separate process. If the deploy utility is used, then setting the elements <deployHarvestDatabaseDir>harvestDatabase</deployHarvestDatabaseDir> and <deployArchiveDatabaseDir>adminDB</deployArchiveDatabaseDir> will automatically result in the deployment and start of databases in the specified directories. If you prefer to configure the databases by hand you should

1. Start Derby separately:
   - cd "directory with the extracted database" (e.g. <deployInstallDir>/<deployHarvestDatabaseDir>)
3. export CLASSPATH=<deployInstallDir>/lib/db/derbynet-10.4.2.0.jar:<deployInstallDir>/lib/db/derby-10.4.2.0.jar
4. java org.apache.derby.drda.NetworkServerControl start -p port

The default port is 1527. Similarly set up a derby for the admin database on its own port.

For the NetarchiveSuite to use this kind of external database, you need to:

- Set the setting `settings.common.database.class` to `dk.netarkivet.harvester.datamodel.DerbyServerSpecifics`.
- Set the setting `settings.common.database.url` to `jdbc:derby://<deployMachine>:1527/fullhddb` (substitute the server host for `<deployMachine>` and 1527 for correct port).

You will need to add a permission to the policy file used by your installation, if you use security (see below). The following will allow NetarchiveSuite to access a Derby database on port 1527.

```java
grant {
    permission java.net.SocketPermission "127.0.0.1:1527",
    "connect, resolve",
};
```

**Firewall note:** You will need to allow the GUIApplication and the HarvestTemplateApplication to be able to access port 1527 on the server where you run the database.

More details on using Derby as a server are available on [http://db.apache.org/derby/docs/dev/adminguide/cadminov825266.html](http://db.apache.org/derby/docs/dev/adminguide/cadminov825266.html) the derby pages.

### PostgreSQL Database

NetarchiveSuite comes with scripts to initialise postgresql databases for both the harvest database and the admin database. These are in:

```
scripts/postgresql/createHarvestDB.pgsql
scripts/postgresql/createAdminDB.pgsql
```

Read the header of `createHarvestDB.pgsql` carefully. It describes how to create a separate tablespace for indexes. If these instructions are not followed, no such tablespace will be created and this will have a deleterious effect on NetarchiveSuites performance.

Also read the header of `createAdminDB.pgsql` for information on how to install the admin database.

The settings for the two database connections look something like (here the harvest database is named 'harvestdatabase', and the admin database is named 'admindb')

```
<settings>
  <archive>
    <admin>
      <class>dk.netarkivet.archive.arcrepositoryadmin.DatabaseAdmin</class>
    </admin>
    <database>
      <class>dk.netarkivet.archive.arcrepositoryadmin.PostgreSQLSpecifics</class>
      <baseUrl>jdbc:postgresql</baseUrl>
      <machine>localhost</machine>
      <port>5432</port>
      <dir>admindb</dir>
      <username>netarchivesuite</username>
      <password>netarchivesuite</password>
    </database>
  </archive>
  <common>
    <database>
      <class>dk.netarkivet.harvester.datamodel.PostgreSQLSpecifics</class>
      <baseUrl>jdbc:postgresql</baseUrl>
      <machine>localhost</machine>
      <port>5432</port>
      <dir>harvestdb</dir>
      <username>netarchivesuite</username>
      <password>netarchivesuite</password>
    </database>
  </common>
</settings>
```
The meaning of the various settings should be fairly obvious: they specify the machine and port-number of the postgresql server, the names of the two databases and the credentials for accessing the databases. In this scenario, you are responsible yourself for configuring postgresql with the necessary databases and users (with read and write permissions to the two databases), and also for initialising the two databases from the supplied schemas.

A sample deploy configuration exists for using postgresql\(\text{deploy\_standalone\_example\_postgresql.xml}\) similar to the one using derby: \text{deploy\_standalone\_example.xml}.

**Choose a JMS broker**

NetarchiveSuite requires a JMS broker to run. The only type of JMS broker supported at this time is the SunMQ broker and its open source counterpart Open Message Queue.

The installation and start-up of a JMS broker is described in \text{Appendix A - Necessary external software}.

For description of how to configure the JMS broker, please refer to the \text{JMS section in Installation Manual}.

**Firewall note**: The machine that runs the JMS broker must be accessible from all machines in the installation on not only port 7676, but also port 33700 (from RMI).

**Java**

All machines must run Java version 1.8.0 or higher (The software is tested with Oracle JDK 1.8.0_20).

**Choose the set of machines taking part in the installation/deployment**

When you have chosen a scenario, you must decide on the number of machines, you want to use in the deployment of the NetarchiveSuite. For scenario A, the answer is of course one. For the scenarios B, C, and D, the answer is more complicated.

An extra complication is added by installing the system at two different physical location (here referred as EAST and WEST). The distinction between different physical location are relevant if the system is installed at two different institutions with firewalls between them.

At the Danish installation, we operate with 4 kinds of machines:

- Admin machine (one server): Here we deploy one or more BitarchiveMonitorApplications (one for each bitarchive Replica), one ArcrepositoryApplication, one GUIApplication, and a JobManagerApplication, which takes care of job scheduling.
- Harvester machines (several at each physical location): Here we deploy the HarvestControllerApplications.
- Bitarchive machines (one or more at each physical location): These machines one or more BitarchiveApplications each (there must be at least one for each bitarchive Replica).
- Access servers (one or more): On these machines, we have the ViewerproxyApplication enabling us to browse in already stored webpages, and the unique IndexServerApplication.

Apart from the HarvestControllerApplications, there is no requirement that the applications are placed like this, but we will use it as an example throughout the rest of the manual. In the standard set-up used in our test-environment, we have 8 machines:

- 1 bitarchive server (on physical location WEST)
- 2 bitarchive servers (on physical location EAST)
- 1 admin machine (placed on physical location EAST)
- 1 harvester-machine (placed on physical location WEST)
- 2 harvester-machines (placed on physical location EAST)
- 1 access server (placed on physical location EAST)