Stacktraces extraction

The database dump contains information about errors encountered by people when using Eclipse. It is composed of several mongodb tables and uses the bson format. Only two tables contain stack traces: `problems` and `incidents`.

Cleaning

The bson files can be read using the bsondump utility, provided with the mongodb client package (mongodb-clients on Debian).

```
bsondump problems.bson --type json > problems.json
```

After conversion the two files are quite big: 37GB for incidents and 2.1 GB for problems.

Unfortunately the utility adds some progress information in the UI that needs to be removed from the output:

```
grep -v 'Progress: ' problems.json > problems_clean.json
```

We also had to remove a few (approx. a dozen of) lines because they embed unparseable source code, characters or asian/binary/utf8/16/256 text. The script tries to JSON-decode all lines one by one, and on failure simply goes to the next line.

For `problems` (the file is reasonably small) the script generates for each line a separate JSON file with only information related to that line. The script for problems extraction is `parse_json_problems.pl`. Output is 820MB and processing time is roughly 45mn.

For `incidents` (file is 37GB) the script generates for each line a separate JSON file with only information related to that line. For the records, trying to generate a single file requires at least twice the size of the file in RAM/SWAP (i.e. roughly 74GB). There are 2084328 files in the output for 17GB. The script for incidents extraction is `parse_json_incidents.pl`. To get an idea of the resources required to process that, the final incidents extraction took roughly 16h on a quite powerful box.

Privacy concerns

The result contains no email address, user id or machine id. Rather than removing the information (we are not sure that we remove all required information) we decided to simply pick relevant information from the file and push it into the output.

End users have an option to keep their own class names private. We have presently no simple means to know what stacktraces in the database extraction should be kept private, so we decided to play it safe and hide class names whose packages don't start with known prefixes [1]. All private classnames have been replaced by the HIDDEN keyword.

[1] "ch.qos.*", "com.cforcoding.*", "com.google.*", "com.gradleware.tooling.*", "com.mountainminds.eclemma.*", "com.naef.*", "com.sun.*", "java.*", "javafx.*", "javax.*", "org.apache.*", "org.eclipse.*", "org.fordiac.*", "org.gradle.*", "org.jacoco.*", "org.osgi.*", "org.slf4j.*", "sun.*"

Format: problems

```
{
  "summary": "",
  "osgiArch": "",
  "osgiOs": "",
  "osgiOsVersion": "",
  "osgiWs": "",
  "eclipseBuildId": "",
  "eclipseProduct": "",
  "javaRuntimeVersion": "",
  "numberOfIncidents": 0,
  "numberOfReporters": 74,
  "stacktraces": [
    [ "stacktrace for incident" ],
    [ "stacktrace for cause" ],
    [ "stacktrace for exception" ]
  ]
}
```

Format: incidents

```
{
  "eclipseBuildId":"4.6.1.M20160907-1200",
  "eclipseProduct":"org.eclipse.epp.package.jee.product",
  "javaRuntimeVersion":"1.8.0_112-b15",
  "osgiArch":"x86_64",
  "osgiOs":"Windows7",
  "osgiOsVersion":"6.1.0",
  "osgiWs":"win32",
  "stacktraces":[
    [ "stacktrace" ]
  ],
  "summary": "Failed to retrieve default libraries for jre1.8.0_111"
}
```

Format: Stacktraces

The structure used in the mongodb for stacktraces has been kept as is: it is composed of fields with all information relevant to each line of the stacktrace. Each stacktrace is an array of objects as shown below:

```
[
  {
    "cN": "sun.net.www.http.HttpClient",
    "mN": "parseHTTPHeader",
    "fN": "HttpClient.java",
    "lN": 786,
  }
]
```