ErrorMessagesInNutch2

Error Messages in Nutch 2.0

This page acts as a repository for potential error messages you might experience whilst using Nutch 2.0. It will most likely be dynamic and fairly general in nature due to the variety of additional software projects which can be combined with Nutch 2.0 and the potential for errors which this presents both for Nutch and which need to be considered when working with other software projects in combination.

Nutch logging shows Skipping http://myurlForParsing.com; different batch id (null)

If your logging level is set to DEBUG, this may occur whilst executing FetcherJob, ParserJob and **IndexerJob**. For example, within ParserJob#map, the situation arises where the NutchJob.shouldProcess returns true due to the fact that Mark.FETCH_MARK.checkMark(page) returns value null. The code for this can be seen below.

[code]

```
@Override
public void map(String key, [WebPage] page, Context context)
throws IOException, [InterruptedException] {
Utf8 mark = Mark.FETCH_MARK.checkMark(page);
String unreverseKey = [TableUtil].unreverseUrl(key);
if (batchId.equals(REPARSE)) {
LOG.debug("Reparsing " + unreverseKey);
} else {
if (NutchJob.shouldProcess(mark, batchId)) {
if (LOG.isDebugEnabled()) {
LOG.debug("Skipping " + [TableUtil].unreverseUrl(key) + "; different batch id (" + mark + ")");
}
return;
```

[code]

What we wish to know is in which scenarios it is possible to have a page which we attempt to fetch, parse or index which has a null value for *_MARK?

- Well, whilst the Jobs are executing this can occur for example as you have to load all backend entries, as there are no filters ("where" clauses in SQL) in Apache Gora. This means that you will see a lot of entries with wrong mark's.
- Null values are possible, too, think about these steps: inject -> generate -> inject -> fetch. The second inject will leave entries in the db without fetchmarks seen by the fetcher later.

It seems to be that updating the web database with the DBUpdaterJob, sorts this out.

gora-cassandra >0.2 InvalidRequestException(why:Keyspace webpage does not exist)

This seems to be encountered when attempting to inject URLs into Cassandra after the server is started and stopped intermittently many times. This may possibly lead to the particular 'webpage' Keyspace and/or data in the Cassandra data directory becoming corrupted. So far, the only solution seems to be deleting the cassandra data directory and starting again. It should be noted that this is not a common error to encounter.

```
2013-02-10 16:32:23,796 WARN  mapred.LocalJobRunner - job_local_0001
me.prettyprint.hector.api.exceptions.HInvalidRequestException: InvalidRequestException(why:Keyspace webpage
does not exist)
        at me.prettyprint.cassandra.connection.client.HThriftClient.getCassandra(HThriftClient.java:80)
        at me.prettyprint.cassandra.connection.HConnectionManager.operateWithFailover(HConnectionManager.java:
251)
        at me.prettyprint.cassandra.model.ExecutingKeyspace.doExecuteOperation(ExecutingKeyspace.java:97)
        at me.prettyprint.cassandra.model.MutatorImpl.execute(MutatorImpl.java:243)
        at me.prettyprint.cassandra.model.MutatorImpl.insert(MutatorImpl.java:69)
        at org.apache.gora.cassandra.store.HectorUtils.insertColumn(HectorUtils.java:47)
        at org.apache.gora.cassandra.store.CassandraClient.addColumn(CassandraClient.java:169)
        at org.apache.gora.cassandra.store.CassandraStore.addOrUpdateField(CassandraStore.java:341)
        at org.apache.gora.cassandra.store.CassandraStore.flush(CassandraStore.java:228)
        at org.apache.gora.cassandra.store.CassandraStore.close(CassandraStore.java:95)
        at org.apache.gora.mapreduce.GoraRecordWriter.close(GoraRecordWriter.java:55)
        at org.apache.hadoop.mapred.MapTask$NewDirectOutputCollector.close(MapTask.java:651)
        at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:766)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:370)
        at org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:212)
Caused by: InvalidRequestException(why:Keyspace webpage does not exist)
        at org.apache.cassandra.thrift.Cassandra$set_keyspace_result.read(Cassandra.java:4874)
        at org.apache.thrift.TServiceClient.receiveBase(TServiceClient.java:78)
        at org.apache.cassandra.thrift.Cassandra$Client.recv_set_keyspace(Cassandra.java:489)
        at org.apache.cassandra.thrift.Cassandra$Client.set_keyspace(Cassandra.java:476)
        at me.prettyprint.cassandra.connection.client.HThriftClient.getCassandra(HThriftClient.java:78)
        ... 14 more
2013-02-10 16:32:24,149 ERROR crawl.InjectorJob - InjectorJob: java.lang.RuntimeException: job failed:
name=inject urls, jobid=job_local_0001
        at org.apache.nutch.util.NutchJob.waitForCompletion(NutchJob.java:54)
        at org.apache.nutch.crawl.InjectorJob.run(InjectorJob.java:233)
        at org.apache.nutch.crawl.InjectorJob.inject(InjectorJob.java:251)
        at org.apache.nutch.crawl.InjectorJob.run(InjectorJob.java:273)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:65)
        at org.apache.nutch.crawl.InjectorJob.main(InjectorJob.java:282)
```

Nutch 2.1 + HBase 0.90.4 cluster settings - WARN zookeeper.ClientCnxn - Session 0x0 for server node1.xxxxxx.com/xxx.xxx.xxx.xxx:2181, unexpected error, closing socket connection and attempting reconnect java.io.IOException: Connection reset by peer

There seems to be a bug in the HBase 0.90.4 library, which comes with Nutch 2.x. If you replace hbase-0.90.4.jar with hbase-0.90.6-cdh3u5.jar (assuming that you are running with CDH3 Update 5) the problem should be resolved and you will avoid the rather nasty looking trace as below. Here is the mailing list thread for reference. **N.B.** Please also ensure that zoo.cfg is copied to the hadoop conf directory (e.g. on classpath) on the entire cluster.

```
2013-02-06 17:09:32,775 WARN  zookeeper.ClientCnxn - Session 0x0 for
server node1.xxxxxx.com/xxx.xxx.xxx.xxx:2181, unexpected error,
closing socket connection and attempting reconnect
java.io.IOException: Connection reset by peer
        at sun.nio.ch.FileDispatcher.read0(Native Method)
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:21)
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:198)
        at sun.nio.ch.IOUtil.read(IOUtil.java:166)
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:245)
        at org.apache.zookeeper.ClientCnxn$SendThread.doIO(ClientCnxn.java:817)
        at org.apache.zookeeper.ClientCnxn$SendThread.run(ClientCnxn.java:1089)
2013-02-06 17:09:34,337 WARN  mapred.FileOutputCommitter - Output path
is null in cleanup
2013-02-06 17:09:34,337 WARN  mapred.LocalJobRunner - job_local_0001
org.apache.hadoop.hbase.ZooKeeperConnectionException: HBase is able to
connect to ZooKeeper but the connection closes immediately. This could
be a sign that the server has too many connections (30 is the
default). Consider inspecting your ZK server logs for that error and
then make sure you are reusing HBaseConfiguration as often as you can.
See HTable's javadoc for more information.
        at
org.apache.hadoop.hbase.zookeeper.ZooKeeperWatcher.<init>(ZooKeeperWatcher.java:155)
        at
org.apache.hadoop.hbase.client.HConnectionManager$HConnectionImplementation.getZooKeeperWatcher
```

```
(HConnectionManager.java:1
002)
        at
org.apache.hadoop.hbase.client.HConnectionManager$HConnectionImplementation.setupZookeeperTrackers
(HConnectionManager.jav
a:304)
        at
org.apache.hadoop.hbase.client.HConnectionManager$HConnectionImplementation.<init>(HConnectionManager.java:295)
        at
org.apache.hadoop.hbase.client.HConnectionManager.getConnection(HConnectionManager.java:157)
        at org.apache.hadoop.hbase.client.HBaseAdmin.<init>(HBaseAdmin.java:90)
        at
org.apache.gora.hbase.store.HBaseStore.initialize(HBaseStore.java:108)
        at
org.apache.gora.store.impl.DataStoreBase.readFields(DataStoreBase.java:181)
        at org.apache.gora.query.impl.QueryBase.readFields(QueryBase.java:222)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:67)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:40)
        at org.apache.gora.util.IOUtils.deserialize(IOUtils.java:217)
        at org.apache.gora.util.IOUtils.deserialize(IOUtils.java:237)
        at
org.apache.gora.query.impl.PartitionQueryImpl.readFields(PartitionQueryImpl.java:141)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:67)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:40)
        at org.apache.gora.util.IOUtils.deserialize(IOUtils.java:217)
        at org.apache.gora.util.IOUtils.deserialize(IOUtils.java:237)
        at
org.apache.gora.mapreduce.GoraInputSplit.readFields(GoraInputSplit.java:76)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:67)
        at
org.apache.hadoop.io.serializer.WritableSerialization$WritableDeserializer.deserialize(WritableSerialization.
java:40)
        at org.apache.hadoop.mapred.MapTask.getSplitDetails(MapTask.java:396)
        at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:728)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:370)
        at
org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:212)
Caused by: org.apache.zookeeper.KeeperException$ConnectionLossException:
KeeperErrorCode = ConnectionLoss for /hbase
        at org.apache.zookeeper.KeeperException.create(KeeperException.java:90)
        at org.apache.zookeeper.KeeperException.create(KeeperException.java:42)
        at org.apache.zookeeper.ZooKeeper.exists(ZooKeeper.java:809)
        at org.apache.zookeeper.ZooKeeper.exists(ZooKeeper.java:837)
        at
org.apache.hadoop.hbase.zookeeper.ZKUtil.createAndFailSilent(ZKUtil.java:903)
        at
org.apache.hadoop.hbase.zookeeper.ZooKeeperWatcher.<init>(ZooKeeperWatcher.java:133)
        ... 24 more
```

Nutch 2.0 and HBase 0.90.4 - org.apache.hadoop.hbase.MasterNotRunningException: master: 60000

The following Exception can occur if the file */etc/hosts* is not configured properly: you have to configure it as described in [0]. For each machine of your cluster, comment the line *127.0.0.1 localhost* and add *localhost* to the line where the master's address is written.

[0] http://stackoverflow.com/questions/7791788/hbase-client-do-not-able-to-connect-with-remote-hbase-server

```
> org.apache.gora.util.GoraException:
> org.apache.hadoop.hbase.MasterNotRunningException: master:60000
>         at
> org.apache.gora.store.DataStoreFactory.createDataStore(DataStoreFactory.java:167)
>         at
> org.apache.gora.store.DataStoreFactory.createDataStore(DataStoreFactory.java:118)
>         at
> org.apache.gora.mapreduce.GoraOutputFormat.getRecordWriter(GoraOutputFormat.java:88)
>         at
> org.apache.hadoop.mapred.MapTask$NewDirectOutputCollector.<init>(MapTask.java:628)
>         at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:753)
>         at org.apache.hadoop.mapred.MapTask.run(MapTask.java:370)
>         at org.apache.hadoop.mapred.Child$4.run(Child.java:255)
>         at java.security.AccessController.doPrivileged(Native Method)
>         at javax.security.auth.Subject.doAs(Subject.java:396)
>         at
> org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1121)
>         at org.apache.hadoop.mapred.Child.main(Child.java:249)
> Caused by: org.apache.hadoop.hbase.MasterNotRunningException: master:60000
>         at
> org.apache.hadoop.hbase.client.HConnectionManager$HConnectionImplementation.getMaster(HConnectionManager.java:
396)
>         at
> org.apache.hadoop.hbase.client.HBaseAdmin.<init>(HBaseAdmin.java:94)
>         at
> org.apache.gora.hbase.store.HBaseStore.initialize(HBaseStore.java:108)
>         at
> org.apache.gora.store.DataStoreFactory.initializeDataStore(DataStoreFactory.java:102)
>         at
> org.apache.gora.store.DataStoreFactory.createDataStore(DataStoreFactory.java:161)
>         ... 10 more
```

Nutch 2.0 and Apache Cassandra

When trying to configure Nutch (running in distributed mode on Cloudera's CDH3) with Cassandra as the Gora storage mechanism, the following NoSuchMethodError results when attempting to inject the crawldb with a seed list.

```
Caused by: java.lang.NoSuchMethodError:
org.apache.thrift.meta_data.FieldValueMetaData.<init>(BZ)V
        at org.apache.cassandra.thrift.CfDef.<clinit>(CfDef.java:299)
        at org.apache.cassandra.thrift.KsDef.read(KsDef.java:753)
        at
org.apache.cassandra.thrift.Cassandra$describe_keyspace_result.read(Cassandra.java:24338)
        at
org.apache.cassandra.thrift.Cassandra$Client.recv_describe_keyspace(Cassandra.java:1371)
        at
org.apache.cassandra.thrift.Cassandra$Client.describe_keyspace(Cassandra.java:1346)
        at
me.prettyprint.cassandra.service.AbstractCluster$4.execute(AbstractCluster.java:192)
        at
me.prettyprint.cassandra.service.AbstractCluster$4.execute(AbstractCluster.java:187)
        at
me.prettyprint.cassandra.service.Operation.executeAndSetResult(Operation.java:101)
        at
me.prettyprint.cassandra.connection.HConnectionManager.operateWithFailover(HConnectionManager.java:232)
        at
me.prettyprint.cassandra.service.AbstractCluster.describeKeyspace(AbstractCluster.java:201)
        at
org.apache.gora.cassandra.store.CassandraClient.checkKeyspace(CassandraClient.java:82)
        at
org.apache.gora.cassandra.store.CassandraClient.init(CassandraClient.java:69)
        at
org.apache.gora.cassandra.store.CassandraStore.<init>(CassandraStore.java:68)
        ... 18 more
```

When using different Gora storage mechanisms we have to manually tweak the Nutch Ivy configuration depending on the choice of Gora store, in this case Cassandra.

To resolve this error the following was added to $NUTCH_HOME/ivy/ivy.xml:

```
<dependency org="org.apache.gora" name="gora-cassandra" rev="0.2-incubating" conf="*->compile"/>
<dependency org="org.apache.cassandra" name="cassandra-thrift" rev="0.8.1"/>
<dependency org="com.ecyrd.speed4j" name="speed4j" rev="0.9" conf="*->*,!javadoc,!sources"/>
<dependency org="com.github.stephenc.high-scale-lib" name="high-scale-lib" rev="1.1.2" conf="*->*,!javadoc,!
sources"/>
<dependency org="com.google.collections" name="google-collections" rev="1.0" conf="*->*,!javadoc,!sources"/>
<dependency org="com.google.guava" name="guava" rev="r09" conf="*->*,!javadoc,!sources"/>
<dependency org="org.apache.cassandra" name="apache-cassandra" rev="0.8.1"/>
<dependency org="me.prettyprint" name="hector-core" rev="0.8.0-2"/>
```

then the following ant commands were executed

```
$ ant clean
$ ant
```

This specified the correct dependencies to be downloaded by Ivy which were then bundled into the nutch-2.0-dev.job file.

In this particular case it was mentioned that Cloudera CDH3 was being used. It has a hue plugins jar with an older thrift library in it, therefore removing this jar from the classpath resolved further errors with running Nutch in distributed mode.

Correspondence on this error can be seen in context here

Missing plugins whilst running Nutch 2.0 on Cloudera's CDH3

Cloudera's CDH3 is Cloudera's distribution including Apache Hadoop. More information can be found here. This common error results due to a bug in MAPREDUCE-967 which modifies the way MapReduce unpacks the job's jar. The old way was to unpack the whole of it, now only classes/ and lib/ are unpacked. This way Nutch is missing the plugins/ directory. A workaround is to force unpacking of the plugin/ directory. If you install only the CDH3 distro, you are OK. It is when you add the Hue distros or try to use a Hadoop installed with the Cloudera SCM products that you run into problems. This can be done by adding the following properties to nutch-site.xml

```
<property>
<name>mapreduce.job.jar.unpack.pattern</name>
<value>(?:classes/|lib/|plugins/).*</value>
</property>

<property>
<name>plugin.folders</name>
<value>${job.local.dir}/../jars/plugins</value>
</property>
```

and by removing hue-plugins-1.2.0-cdh3u1.jar from the hadoop lib folder (e.g. /usr/lib/hadoop-0.20/lib).

It is then necessary to recreate the Nutch job file using ant. Then finally it is important to set HADOOP_OPTS="-Djob.local.dir=/<MY HOME>/nutch /plugins" in hadoop-env.sh.

Although this is a real nasty workaround it does work.

java.lang.RuntimeException compile failure with Gora trunk (1153872)

Although this problem is not specifically related to Nutch 2.0, it still prevents us from compiling the complete Gora infrastructure required to facilitate Nutch activities.

Upon checking out Gora trunk (1153872) and compiling the code you may get the rather nasty runtime exception as follows:

```
BUILD FAILED
/home/lewis/ASF/gora/build.xml:272: The following error occurred while executing this line:
/home/lewis/ASF/gora/build-common.xml:350: impossible to ivy retrieve: java.lang.RuntimeException: problem
during retrieve of org.apache.gora#gora-cassandra: java.lang.RuntimeException: Multiple artifacts of the module
org.apache.cassandra#cassandra-thrift;0.8.1 are retrieved to the same file! Update the retrieve pattern  to fix
this error.
    at org.apache.ivy.core.retrieve.RetrieveEngine.retrieve(RetrieveEngine.java:206)
    at org.apache.ivy.Ivy.retrieve(Ivy.java:540)
    at org.apache.ivy.ant.IvyRetrieve.doExecute(IvyRetrieve.java:59)
    at org.apache.ivy.ant.IvyTask.execute(IvyTask.java:277)
    at org.apache.tools.ant.UnknownElement.execute(UnknownElement.java:291)
    at sun.reflect.GeneratedMethodAccessor6.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:616)
    at org.apache.tools.ant.dispatch.DispatchUtils.execute(DispatchUtils.java:106)
    at org.apache.tools.ant.Task.perform(Task.java:348)
    at org.apache.tools.ant.Target.execute(Target.java:390)
    at org.apache.tools.ant.Target.performTasks(Target.java:411)
    at org.apache.tools.ant.Project.executeSortedTargets(Project.java:1397)
    at org.apache.tools.ant.helper.SingleCheckExecutor.executeTargets(SingleCheckExecutor.java:38)
    at org.apache.tools.ant.Project.executeTargets(Project.java:1249)
    at org.apache.tools.ant.taskdefs.Ant.execute(Ant.java:442)
    at org.apache.tools.ant.taskdefs.SubAnt.execute(SubAnt.java:302)
    at org.apache.tools.ant.taskdefs.SubAnt.execute(SubAnt.java:221)
    at org.apache.tools.ant.UnknownElement.execute(UnknownElement.java:291)
    at sun.reflect.GeneratedMethodAccessor6.invoke(Unknown Source)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:616)
    at org.apache.tools.ant.dispatch.DispatchUtils.execute(DispatchUtils.java:106)
    at org.apache.tools.ant.Task.perform(Task.java:348)
    at org.apache.tools.ant.Target.execute(Target.java:390)
    at org.apache.tools.ant.Target.performTasks(Target.java:411)
    at org.apache.tools.ant.Project.executeSortedTargets(Project.java:1397)
    at org.apache.tools.ant.Project.executeTarget(Project.java:1366)
    at org.apache.tools.ant.helper.DefaultExecutor.executeTargets(DefaultExecutor.java:41)
    at org.apache.tools.ant.Project.executeTargets(Project.java:1249)
    at org.apache.tools.ant.Main.runBuild(Main.java:801)
    at org.apache.tools.ant.Main.startAnt(Main.java:218)
    at org.apache.tools.ant.launch.Launcher.run(Launcher.java:280)
    at org.apache.tools.ant.launch.Launcher.main(Launcher.java:109)
Caused by: java.lang.RuntimeException: Multiple artifacts of the module org.apache.cassandra#cassandra-thrift;
0.8.1 are retrieved to the same file! Update the retrieve pattern  to fix this error.
    at org.apache.ivy.core.retrieve.RetrieveEngine.determineArtifactsToCopy(RetrieveEngine.java:360)
    at org.apache.ivy.core.retrieve.RetrieveEngine.retrieve(RetrieveEngine.java:104)
    ... 33 more
```

This may be due to one other project that has already written to the same file, but with a different revision. The ivy retrive directory and configuration is defined at build.xml as:

```
 <property name="ivy.local.default.root"
value="${ivy.default.ivy.user.dir}/local" />
 <property name="ivy.local.default.ivy.pattern"
value="[organisation]/[module]/[revision]/[type]s/[artifact].[ext]" />
 <property name="ivy.local.default.artifact.pattern"
value="[organisation]/[module]/[revision]/[type]s/[artifact].[ext]" />

 <property name="ivy.shared.default.root"
value="${ivy.default.ivy.user.dir}/shared" />
 <property name="ivy.shared.default.ivy.pattern"
value="[organisation]/[module]/[revision]/[type]s/[artifact].[ext]" />
 <property name="ivy.shared.default.artifact.pattern"
value="[organisation]/[module]/[revision]/[type]s/[artifact].[ext]" />
```

This means that Gora uses the shared default root directory which is
~/.ivy2/

Therefore applying the following patch:

```
diff --git a/build.xml b/build.xml
index 5810db9..eaa0450 100644
--- a/build.xml
+++ b/build.xml
@@ -301,7 +301,7 @@
  </target>

  <!-- target: clean-cache
================================================== -->
-  <target name="clean-cache" depends=""
+  <target name="clean-cache" depends="ivy-init"
                       description="delete ivy cache">
    <ivy:cleancache />
  </target>
}}
and running
{{{
$ ant clean-cache.
$ ant
```

should hopefully solve the problem and result in a successful build.

ps. you can alternatively do

```
$rm -rf ~/ivy2/cache
```