Java-Success.com

Prepare to fast-track, choose & go places with 800+ Java & Big Data Q&As with lots of code & diagrams.

search here …   Go

Home   Why? ▾   300+ Java FAQs ▾   300+ Big Data FAQs ▾   Courses ▾

👤 Membership ▾   Your Career ▾

05: Convert XML file To an Avro File with avro-maven-plugin & Apache Spark

📅 Posted on May 21, 2016

This extends 04: Convert XML file To an Avro File with Apache Spark – writing & reading. Instead of using the GenericRecord, let's generate an avro schema object from the avro schema.

Step 1: The pom.xml file should include the **Apache Spark** & **Avro** libraries as shown below. Pay attetion to the **avro-maven-plugin** that generates POJOs from the avro schemas.

300+ Java Interview FAQs

300+ Java FAQs 🔥   ⌄

16+ Java Key Areas Q&As   ⌄

150+ Java Architect FAQs   ⌄

80+ Java Code Quality Q&As   ⌄

150+ Java Coding Q&As   ⌄

300+ Big Data Interview FAQs

300+ Big Data FAQs 🔥   ⌄

Tutorials - Big Data   ›

TUT - 🔢 Starting Big Data

TUT - Starting Spark & Scala

```xml
1
2
3   <project xmlns="http://maven.apache.org/POM/4.0.0
4       xsi:schemaLocation="http://maven.apache.org/
5       <modelVersion>4.0.0</modelVersion>
6       <groupId>com.mytutorial</groupId>
7       <artifactId>sequence-file</artifactId>
8       <packaging>jar</packaging>
9       <version>1.0-SNAPSHOT</version>
10      <name>sequence-file</name>
11      <url>http://maven.apache.org</url>
12
13      <properties>
14          <maven.compiler.source>1.8</maven.compile
15          <maven.compiler.target>1.8</maven.compile
16          <project.build.sourceEncoding>UTF-8</proj
17          <junit.version>4.8.1</junit.version>
18          <hadoop.version>2.7.2</hadoop.version>
19          <spark-version>1.3.0</spark-version>
20          <avro.version>1.7.7</avro.version>
21      </properties>
22
23      <dependencies>
24          <!-- JUnit -->
25          <dependency>
26              <groupId>junit</groupId>
27              <artifactId>junit</artifactId>
28              <version>${junit.version}</version>
29              <scope>test</scope>
30          </dependency>
31
32          <!-- Hadoop -->
33          <dependency>
34              <groupId>org.apache.hadoop</groupId>
35              <artifactId>hadoop-hdfs</artifactId>
36              <version>${hadoop.version}</version>
37              <exclusions>
38                  <exclusion>
39                      <groupId>javax.servlet</group
40                      <artifactId>*</artifactId>
41                  </exclusion>
42              </exclusions>
43          </dependency>
44          <dependency>
45              <groupId>org.apache.hadoop</groupId>
46              <artifactId>hadoop-client</artifactId
47              <version>${hadoop.version}</version>
48              <exclusions>
49                  <exclusion>
50                      <groupId>javax.servlet</group
51                      <artifactId>*</artifactId>
52                  </exclusion>
53              </exclusions>
54          </dependency>
55
```

800+ Java Interview Q&As

300+ Core Java Q&As ⌄

300+ Enterprise Java Q&As ⌄

150+ Java Frameworks Q&As ⌄

120+ Companion Tech Q&As ⌄

Tutorials - Enterprise Java ⌄

```xml
 56            <!-- Apache Spark -->
 57            <dependency>
 58                <groupId>org.apache.spark</groupId>
 59                <artifactId>spark-core_2.11</artifac
 60                <version>${spark-version}</version>
 61                <exclusions>
 62                    <exclusion>
 63                        <groupId>javax.servlet</grou
 64                        <artifactId>*</artifactId>
 65                    </exclusion>
 66                </exclusions>
 67            </dependency>
 68
 69            <!-- AVRO -->
 70            <dependency>
 71                <groupId>org.apache.avro</groupId>
 72                <artifactId>avro</artifactId>
 73                <version>${avro.version}</version>
 74            </dependency>
 75             <dependency>
 76                <groupId>org.apache.avro</groupId>
 77                <artifactId>avro-mapred</artifactId>
 78                <version>${avro.version}</version>
 79                <classifier>hadoop2</classifier>
 80                <exclusions>
 81                    <exclusion>
 82                        <groupId>org.apache.avro</gr
 83                        <artifactId>*</artifactId>
 84                    </exclusion>
 85                </exclusions>
 86            </dependency>
 87        </dependencies>
 88
 89        <build>
 90            <plugins>
 91                <plugin>
 92                    <groupId>org.apache.avro</groupI
 93                    <artifactId>avro-maven-plugin</a
 94                    <version>${avro.version}</versio
 95                    <executions>
 96                        <execution>
 97                            <phase>generate-sources<
 98                            <goals>
 99                                <goal>schema</goal>
100                            </goals>
101                            <configuration>
102                                <includes>
103                                    <include>trans-r
104                                </includes>
105
106                                <sourceDirectory>${p
107                                <outputDirectory>${p
108                            </configuration>
109
110                        </execution>
```

```
111                     </executions>
112                 </plugin>
113             </plugins>
114         <pluginManagement>
115           <plugins>
116             <!--This plugin's configuration is us
117                 only. It has no influence on the
118             <plugin>
119                 <groupId>org.eclipse.m2e</groupI
120                 <artifactId>lifecycle-mapping</a
121                 <version>1.0.0</version>
122                 <configuration>
123                     <lifecycleMappingMetadata>
124                         <pluginExecutions>
125                             <pluginExecution>
126                                 <pluginExecution
127                                     <groupId>
128                                         org.apac
129                                     </groupId>
130                                     <artifactId>
131                                         avro-mav
132                                     </artifactId
133                                     <versionRang
134                                         [1.8.0,)
135                                     </versionRan
136                                     <goals>
137                                         <goal>sc
138                                     </goals>
139                                 </pluginExecutio
140                                 <action>
141                                     <ignore></ig
142                                 </action>
143                             </pluginExecution>
144                         </pluginExecutions>
145                     </lifecycleMappingMetadata>
146                 </configuration>
147             </plugin>
148         </plugins>
149     </pluginManagement>
150
151     </build>
152 </project>
153
```

Step 2: The **report.xml** file under "src/main/resources/data".

```
1
2  <?xml version="1.0" encoding="UTF-8"?>
3  <transactionReports xmlns="http://mytutorial.com/
4      <transactionReport>
```

```
 5        <report>
 6            <reportNumber>9999</reportNumber>
 7            <createdDatetime>2015-06-15T11:29:52+1
 8            <processedDatetime>2015-06-15T11:29:52
 9            <reportStatusCode>Active</reportStatus
10        </report>
11      </transactionReport>
12 </transactionReports>
13
```

Step 3: The avro schema file "**trans-report.avsc**" under "src/main/resources/schema".

```
 1
 2 {"namespace": "mytutorial.com.report",
 3  "type": "record",
 4  "name": "ReportAvro",
 5  "fields": [
 6      {"name": "reportNumber", "type": "string"},
 7      {"name": "createdDatetime", "type": "string"
 8      {"name": "processedDatetime", "type": "string
 9      {"name": "reportStatusCode", "type": "string
10  ]
11 }
12
```

Step 4: Generate the "ReportAvro.java" by running the maven task "generate-sources". You can run this within eclipse by right clicking on the pom.xml and then "Run as" "maven generate-sources" or on the command-line as

```
 1
 2 mvn generate-sources
 3
```

Within eclipse, select "generated-sources" folder right-click and select "Build Path" and then "Use as Source Folder".

Step 5: Finally, the "**ConvertXmlToAvroWithAvroPluginAndSpark.java**"

class that make use of the generated
ReportAvro.java instead of the GenericRecord class.

```
1
2   package com.mytutorial;
3
4   import java.io.File;
5   import java.io.IOException;
6   import java.io.StringReader;
7   import java.net.URL;
8   import java.util.ArrayList;
9   import java.util.Iterator;
10  import java.util.List;
11
12  import javax.xml.namespace.NamespaceContext;
13  import javax.xml.xpath.XPath;
14  import javax.xml.xpath.XPathConstants;
15  import javax.xml.xpath.XPathExpressionException;
16  import javax.xml.xpath.XPathFactory;
17
18  import mytutorial.com.report.ReportAvro;
19
20  import org.apache.avro.mapred.AvroKey;
21  import org.apache.avro.mapreduce.AvroJob;
22  import org.apache.avro.mapreduce.AvroKeyInputForr
23  import org.apache.avro.mapreduce.AvroKeyOutputFoi
24  import org.apache.commons.io.FileUtils;
25  import org.apache.hadoop.conf.Configuration;
26  import org.apache.hadoop.io.NullWritable;
27  import org.apache.hadoop.mapreduce.Job;
28  import org.apache.spark.SparkConf;
29  import org.apache.spark.api.java.JavaPairRDD;
30  import org.apache.spark.api.java.JavaSparkContext
31  import org.w3c.dom.Node;
32  import org.xml.sax.InputSource;
33
34  import scala.Tuple2;
35
36  import com.mytutorial.pojo.Report;
37
38  public class ConvertXmlToAvroWithAvroPluginAndSpa
39
40      private static final String FILE_IN_PATH = "(
41      private static final String FILE_OUT_PATH =
42      private static final String AVRO_SCHEMA_FILE
43
44      final static JavaSparkContext sc;
45
46      static {
47          SparkConf conf = new SparkConf()
48                  .setAppName("Sequence To Avro")
49                  .setMaster("local[1]")
```

```
50              .set("spark.executor.memory", "1g
51              .set("spark.serializer",
52                  "org.apache.spark.serial
53          sc = new JavaSparkContext(conf);
54      }
55
56      public static void main(String[] args) throws
57              XPathExpressionException {
58          URL resource = ConvertXmlToSequenceWithSp
59
60          File inputFile = new File(resource.getPa
61          File outputFile = new File(resource.getP
62          File avroSchemaFile = new File(resource.
63                  + AVRO_SCHEMA_FILE);
64
65          Report report = convertXmlToPojo(inputFi
66
67          write(avroSchemaFile, outputFile, report
68
69          read(outputFile);
70
71          sc.stop();
72      }
73
74      public static void write(File avroSchemaFile
75              throws IOException {
76
77
78          ReportAvro reportAVro= new ReportAvro();
79          reportAVro.setReportNumber(report.getRep
80          reportAVro.setCreatedDatetime(report.getC
81          reportAVro.setProcessedDatetime(report.g
82          reportAVro.setReportStatusCode(report.get
83
84          Job job = Job.getInstance();
85          AvroJob.setOutputKeySchema(job, reportAV
86
87          ArrayList<Tuple2<AvroKey<ReportAvro>, Vo
88          arrayList.add(new Tuple2<AvroKey<ReportAv
89                  reportAVro), null));
90
91          JavaPairRDD<AvroKey<ReportAvro>, Void> h
92                  .parallelizePairs(arrayList);
93
94          if (outputFile.exists()) {
95              FileUtils.deleteQuietly(outputFile);
96          }
97
98          hadoopRDD.saveAsNewAPIHadoopFile(outputF
99                  NullWritable.class, AvroKeyOutpu
100                 job.getConfiguration());
101
102     }
103
104     public static void read(File avroFileToRead)
```

```
105
106          @SuppressWarnings("unchecked")
107          JavaPairRDD<AvroKey<ReportAvro>, NullWri
108                  .newAPIHadoopFile(avroFileToRead
109                          AvroKeyInputFormat.class
110                          NullWritable.class, new
111
112          List<AvroKey<ReportAvro>> avroKeys = avr
113          AvroKey<ReportAvro> avroKey = avroKeys.ge
114
115          System.out.println("Avro Data=" + avroKey
116      }
117
118      // Xpath to read XML & convert it to a pojo
119      private static Report convertXmlToPojo(File
120              XPathExpressionException {
121          XPath xPath = XPathFactory.newInstance()
122          // namespace
123          NamespaceContext ctx = new NamespaceConte
124              public String getNamespaceURI(String
125                  return prefix.equals("urn") ? "h
126                          : null;
127              }
128
129              public Iterator<String> getPrefixes(
130                  return null;
131              }
132
133              public String getPrefix(String uri)
134                  return null;
135              }
136          };
137
138          xPath.setNamespaceContext(ctx);
139          String str = FileUtils.readFileToString(
140          StringReader sr = new StringReader(str);
141          InputSource source = new InputSource(sr)
142
143          // get the DOM
144          Node root = (Node) xPath.evaluate("/", so
145
146          // use the DOM
147          String reportNumber = xPath.evaluate("//
148                  root);
149          String createdDatetime = xPath.evaluate(
150                  "//urn:report/urn:createdDatetim
151          String processedDatetime = xPath.evaluate
152                  "//urn:report/urn:processedDateti
153          String reportStatusCode = xPath.evaluate
154                  "//urn:report/urn:reportStatusCo
155
156          return new Report(reportNumber, createdDa
157                  reportStatusCode);
158
159      }
```

```
160  }
161
```

Output:

```
1
2  Avro Data={"reportNumber": "9999", "createdDatetim
3
```

Note: The generate ReportAvro.java class:

```
1
2   /**
3    * Autogenerated by Avro
4    *
5    * DO NOT EDIT DIRECTLY
6    */
7   package mytutorial.com.report;
8   @SuppressWarnings("all")
9   @org.apache.avro.specific.AvroGenerated
10  public class ReportAvro extends org.apache.avro.:
11    public static final org.apache.avro.Schema SCHE
12    public static org.apache.avro.Schema getClassSe
13    @Deprecated public java.lang.CharSequence repoi
14    @Deprecated public java.lang.CharSequence crea
15    @Deprecated public java.lang.CharSequence proce
16    @Deprecated public java.lang.CharSequence repoi
17
18    /**
19     * Default constructor.  Note that this does no
20     * to their default values from the schema.  I
21     * one should use <code>newBuilder()</code>.
22     */
23    public ReportAvro() {}
24
25    /**
26     * All-args constructor.
27     */
28    public ReportAvro(java.lang.CharSequence repor
29      this.reportNumber = reportNumber;
30      this.createdDatetime = createdDatetime;
31      this.processedDatetime = processedDatetime;
32      this.reportStatusCode = reportStatusCode;
33    }
34
35    public org.apache.avro.Schema getSchema() { re
36    // Used by DatumWriter.  Applications should no
37    public java.lang.Object get(int field$) {
```

```
38      switch (field$) {
39      case 0: return reportNumber;
40      case 1: return createdDatetime;
41      case 2: return processedDatetime;
42      case 3: return reportStatusCode;
43      default: throw new org.apache.avro.AvroRuntir
44      }
45    }
46    // Used by DatumReader.  Applications should no
47    @SuppressWarnings(value="unchecked")
48    public void put(int field$, java.lang.Object va
49      switch (field$) {
50      case 0: reportNumber = (java.lang.CharSequenc
51      case 1: createdDatetime = (java.lang.CharSequ
52      case 2: processedDatetime = (java.lang.CharSe
53      case 3: reportStatusCode = (java.lang.CharSe
54      default: throw new org.apache.avro.AvroRuntir
55      }
56    }
57
58    /**
59     * Gets the value of the 'reportNumber' field.
60     */
61    public java.lang.CharSequence getReportNumber(
62      return reportNumber;
63    }
64
65    /**
66     * Sets the value of the 'reportNumber' field.
67     * @param value the value to set.
68     */
69    public void setReportNumber(java.lang.CharSequ
70      this.reportNumber = value;
71    }
72
73    /**
74     * Gets the value of the 'createdDatetime' fiel
75     */
76    public java.lang.CharSequence getCreatedDatetir
77      return createdDatetime;
78    }
79
80    /**
81     * Sets the value of the 'createdDatetime' fiel
82     * @param value the value to set.
83     */
84    public void setCreatedDatetime(java.lang.CharSe
85      this.createdDatetime = value;
86    }
87
88    /**
89     * Gets the value of the 'processedDatetime' f
90     */
91    public java.lang.CharSequence getProcessedDate
92      return processedDatetime;
```

```
 93      }
 94
 95      /**
 96       * Sets the value of the 'processedDatetime' f·
 97       * @param value the value to set.
 98       */
 99      public void setProcessedDatetime(java.lang.Cha
100        this.processedDatetime = value;
101      }
102
103      /**
104       * Gets the value of the 'reportStatusCode' fi
105       */
106      public java.lang.CharSequence getReportStatusC
107        return reportStatusCode;
108      }
109
110      /**
111       * Sets the value of the 'reportStatusCode' fi
112       * @param value the value to set.
113       */
114      public void setReportStatusCode(java.lang.Char
115        this.reportStatusCode = value;
116      }
117
118      /** Creates a new ReportAvro RecordBuilder */
119      public static mytutorial.com.report.ReportAvro
120        return new mytutorial.com.report.ReportAvro.
121      }
122
123      /** Creates a new ReportAvro RecordBuilder by
124      public static mytutorial.com.report.ReportAvro
125        return new mytutorial.com.report.ReportAvro.
126      }
127
128      /** Creates a new ReportAvro RecordBuilder by
129      public static mytutorial.com.report.ReportAvro
130        return new mytutorial.com.report.ReportAvro.
131      }
132
133      /**
134       * RecordBuilder for ReportAvro instances.
135       */
136      public static class Builder extends org.apache
137        implements org.apache.avro.data.RecordBuilde
138
139        private java.lang.CharSequence reportNumber;
140        private java.lang.CharSequence createdDateti
141        private java.lang.CharSequence processedDate·
142        private java.lang.CharSequence reportStatusC
143
144        /** Creates a new Builder */
145        private Builder() {
146          super(mytutorial.com.report.ReportAvro.SCH
147        }
```

```
148
149      /** Creates a Builder by copying an existing
150      private Builder(mytutorial.com.report.Report
151        super(other);
152        if (isValidValue(fields()[0], other.reportN
153          this.reportNumber = data().deepCopy(field
154          fieldSetFlags()[0] = true;
155        }
156        if (isValidValue(fields()[1], other.created
157          this.createdDatetime = data().deepCopy(f
158          fieldSetFlags()[1] = true;
159        }
160        if (isValidValue(fields()[2], other.process
161          this.processedDatetime = data().deepCopy(
162          fieldSetFlags()[2] = true;
163        }
164        if (isValidValue(fields()[3], other.reportS
165          this.reportStatusCode = data().deepCopy(
166          fieldSetFlags()[3] = true;
167        }
168      }
169
170      /** Creates a Builder by copying an existing
171      private Builder(mytutorial.com.report.Report
172            super(mytutorial.com.report.ReportAv
173        if (isValidValue(fields()[0], other.reportN
174          this.reportNumber = data().deepCopy(field
175          fieldSetFlags()[0] = true;
176        }
177        if (isValidValue(fields()[1], other.created
178          this.createdDatetime = data().deepCopy(f
179          fieldSetFlags()[1] = true;
180        }
181        if (isValidValue(fields()[2], other.process
182          this.processedDatetime = data().deepCopy(
183          fieldSetFlags()[2] = true;
184        }
185        if (isValidValue(fields()[3], other.reportS
186          this.reportStatusCode = data().deepCopy(
187          fieldSetFlags()[3] = true;
188        }
189      }
190
191      /** Gets the value of the 'reportNumber' fie
192      public java.lang.CharSequence getReportNumber
193        return reportNumber;
194      }
195
196      /** Sets the value of the 'reportNumber' fie
197      public mytutorial.com.report.ReportAvro.Build
198        validate(fields()[0], value);
199        this.reportNumber = value;
200        fieldSetFlags()[0] = true;
201        return this;
202      }
```

```
203
204      /** Checks whether the 'reportNumber' field |
205      public boolean hasReportNumber() {
206        return fieldSetFlags()[0];
207      }
208
209      /** Clears the value of the 'reportNumber' f
210      public mytutorial.com.report.ReportAvro.Build
211        reportNumber = null;
212        fieldSetFlags()[0] = false;
213        return this;
214      }
215
216      /** Gets the value of the 'createdDatetime'
217      public java.lang.CharSequence getCreatedDate
218        return createdDatetime;
219      }
220
221      /** Sets the value of the 'createdDatetime'
222      public mytutorial.com.report.ReportAvro.Build
223        validate(fields()[1], value);
224        this.createdDatetime = value;
225        fieldSetFlags()[1] = true;
226        return this;
227      }
228
229      /** Checks whether the 'createdDatetime' fie
230      public boolean hasCreatedDatetime() {
231        return fieldSetFlags()[1];
232      }
233
234      /** Clears the value of the 'createdDatetime
235      public mytutorial.com.report.ReportAvro.Build
236        createdDatetime = null;
237        fieldSetFlags()[1] = false;
238        return this;
239      }
240
241      /** Gets the value of the 'processedDatetime
242      public java.lang.CharSequence getProcessedDa
243        return processedDatetime;
244      }
245
246      /** Sets the value of the 'processedDatetime
247      public mytutorial.com.report.ReportAvro.Build
248        validate(fields()[2], value);
249        this.processedDatetime = value;
250        fieldSetFlags()[2] = true;
251        return this;
252      }
253
254      /** Checks whether the 'processedDatetime' f
255      public boolean hasProcessedDatetime() {
256        return fieldSetFlags()[2];
257      }
```

```
258
259        /** Clears the value of the 'processedDateti
260        public mytutorial.com.report.ReportAvro.Buil
261          processedDatetime = null;
262          fieldSetFlags()[2] = false;
263          return this;
264        }
265
266        /** Gets the value of the 'reportStatusCode'
267        public java.lang.CharSequence getReportStatu
268          return reportStatusCode;
269        }
270
271        /** Sets the value of the 'reportStatusCode'
272        public mytutorial.com.report.ReportAvro.Buil
273          validate(fields()[3], value);
274          this.reportStatusCode = value;
275          fieldSetFlags()[3] = true;
276          return this;
277        }
278
279        /** Checks whether the 'reportStatusCode' fi
280        public boolean hasReportStatusCode() {
281          return fieldSetFlags()[3];
282        }
283
284        /** Clears the value of the 'reportStatusCod
285        public mytutorial.com.report.ReportAvro.Buil
286          reportStatusCode = null;
287          fieldSetFlags()[3] = false;
288          return this;
289        }
290
291        @Override
292        public ReportAvro build() {
293          try {
294            ReportAvro record = new ReportAvro();
295            record.reportNumber = fieldSetFlags()[0]
296            record.createdDatetime = fieldSetFlags()
297            record.processedDatetime = fieldSetFlags
298            record.reportStatusCode = fieldSetFlags(
299            return record;
300          } catch (Exception e) {
301            throw new org.apache.avro.AvroRuntimeExc
302          }
303        }
304      }
305    }
306
```

‹   05: 7 Java FP (lambda expressions) real life examples in wrangling

normal & big data

04: Drools with rules in an excel spreadsheet – Data validation   ›

Disclaimer

The contents in this Java-Success are copyrighted and from EmpoweringTech pty ltd. The EmpoweringTech pty ltd has the right to correct or enhance the current content without any prior notice. These are general advice only, and one needs to take his/her own circumstances into consideration. The EmpoweringTech pty ltd will not be held liable for any damages caused or alleged to be caused either directly or indirectly by these materials and resources. Any trademarked names or labels used in this blog remain the property of their respective trademark owners. Links to external sites do not imply endorsement of the linked-to sites. Privacy Policy

© 2022 java-success.com