医療情報（MML3.0）Hadoopプログラミング

2011年10月21日

キヤノンソフト情報システム
上村準也

発表の経緯

- パッケージ製品の開発者です
  - JavaでWebアプリケーションを開発
  - PaaS/IaaSや仮想マシンを使うことに興味あり
- Hadoopについては
  - 技術的なニュースは度々聞くが使ったことはない
  - 仕事でも案件として話を聞くことがあるように
- 2011年1月にHadoopのハンズオン
  - クラウド・テクノロジー分野第四回勉強会で初めて
  - 本業とは関係なく勉強することに

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

発表の内容

- 医療情報（MML3.0）を題材にHadoopでプログラミング
  - Hadoop Core 0.21.0 を使用
  - MapReduceアルゴリズムの話ではなく…

1. 開発環境の準備
2. テストコードの記述
3. 圧縮データの入力
4. XMLデータの入出力

  - XMLデータをどう扱うのが良い？
  - 象本とHadoop徹底入門を読みながら試したこと

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備(１)

- 開発環境 – Windows7 および Windows XP SP3
- ダウンロード＆インストール
  - JDK – Oracle 1.6.0_27
  - Eclipse – Indigo Service Release 1
  - Gnu On Windows (Gow) – 0.4.0
  - Hadoop – 0.21.0

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備（２）

- Hadoopの設定ファイルを編集

conf/core-site.xml

```
<property>
  <name>hadoop.tmp.dir</name>
  <value>E:¥TEMP¥hadoop</value>
</property>
```

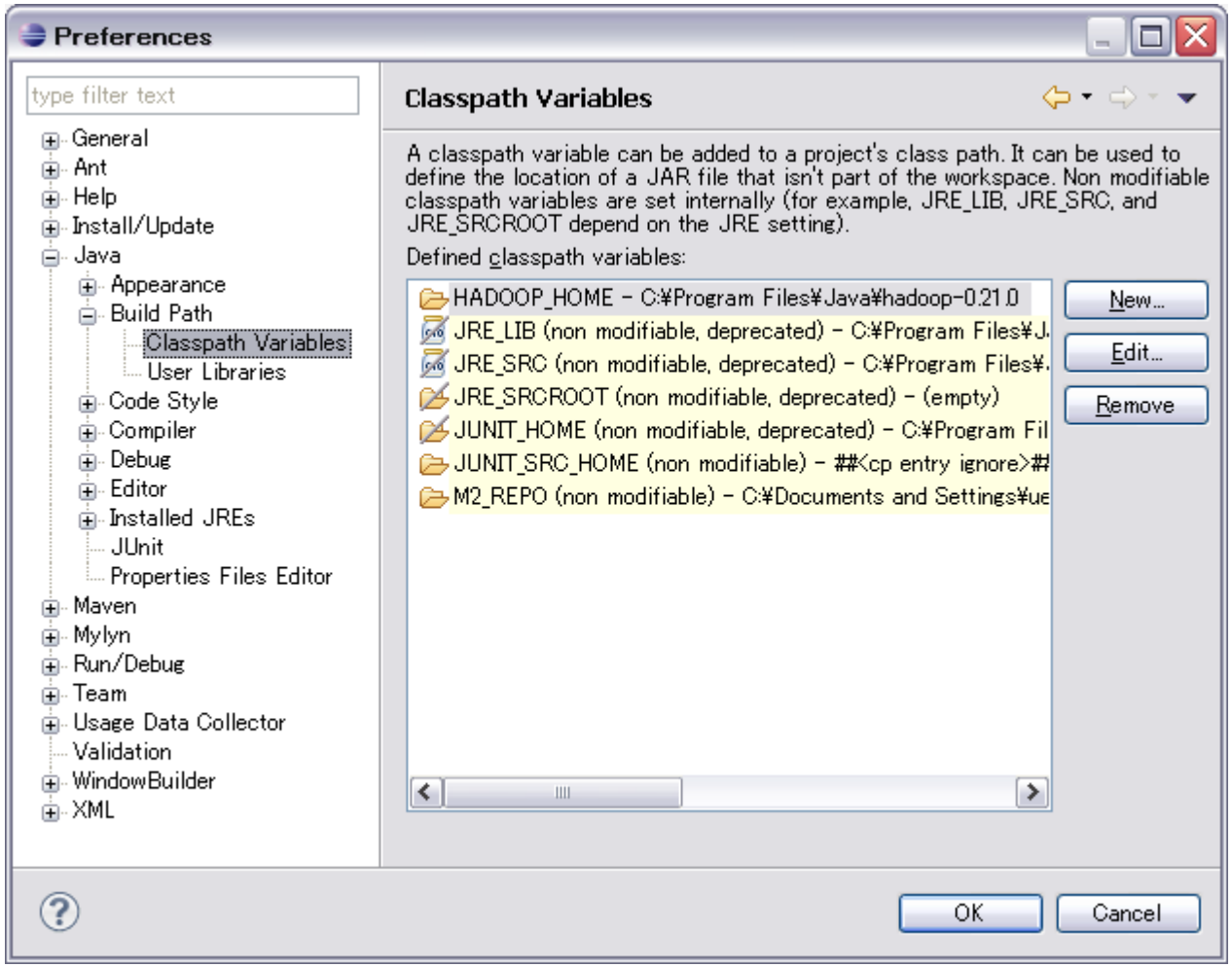
conf/mapred-site.xml

```
<property>
  <name>mapreduce.jobtracker.address</name>
  <value>local</value>
</property>
```

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備(３)

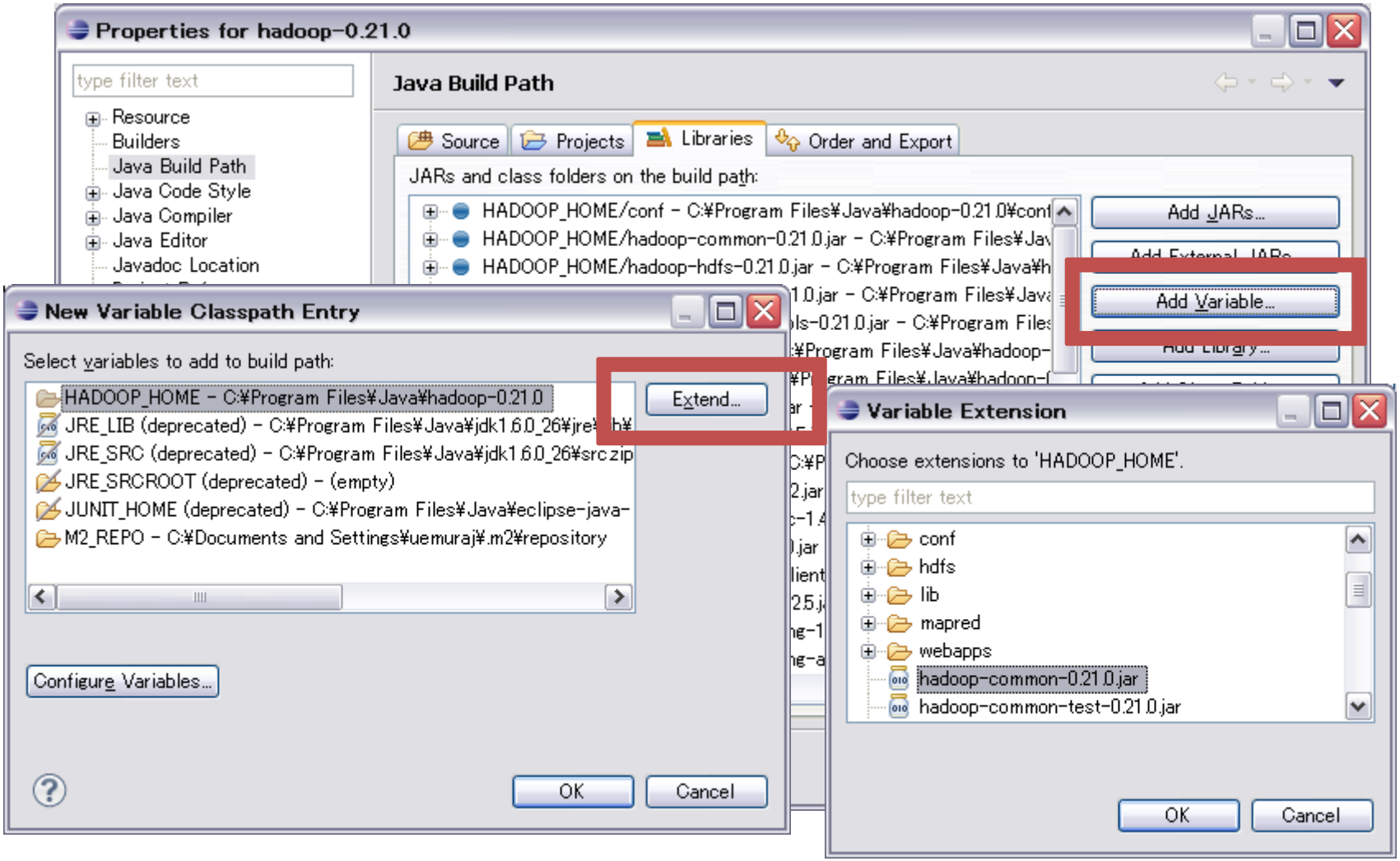
- Eclipseのクラスパス変数を追加



Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備(４)

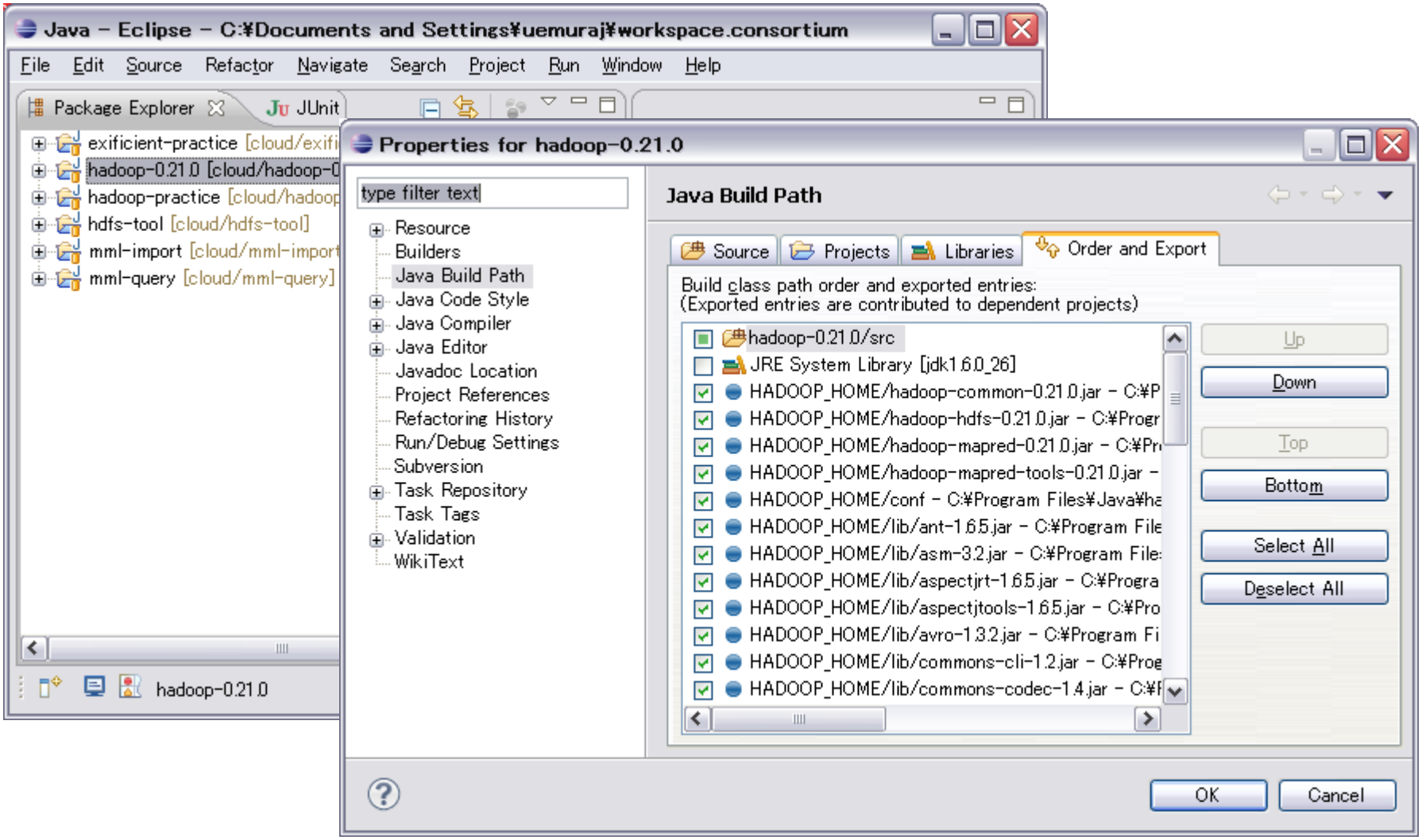
- プロジェクトにライブラリを追加



Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備(５)

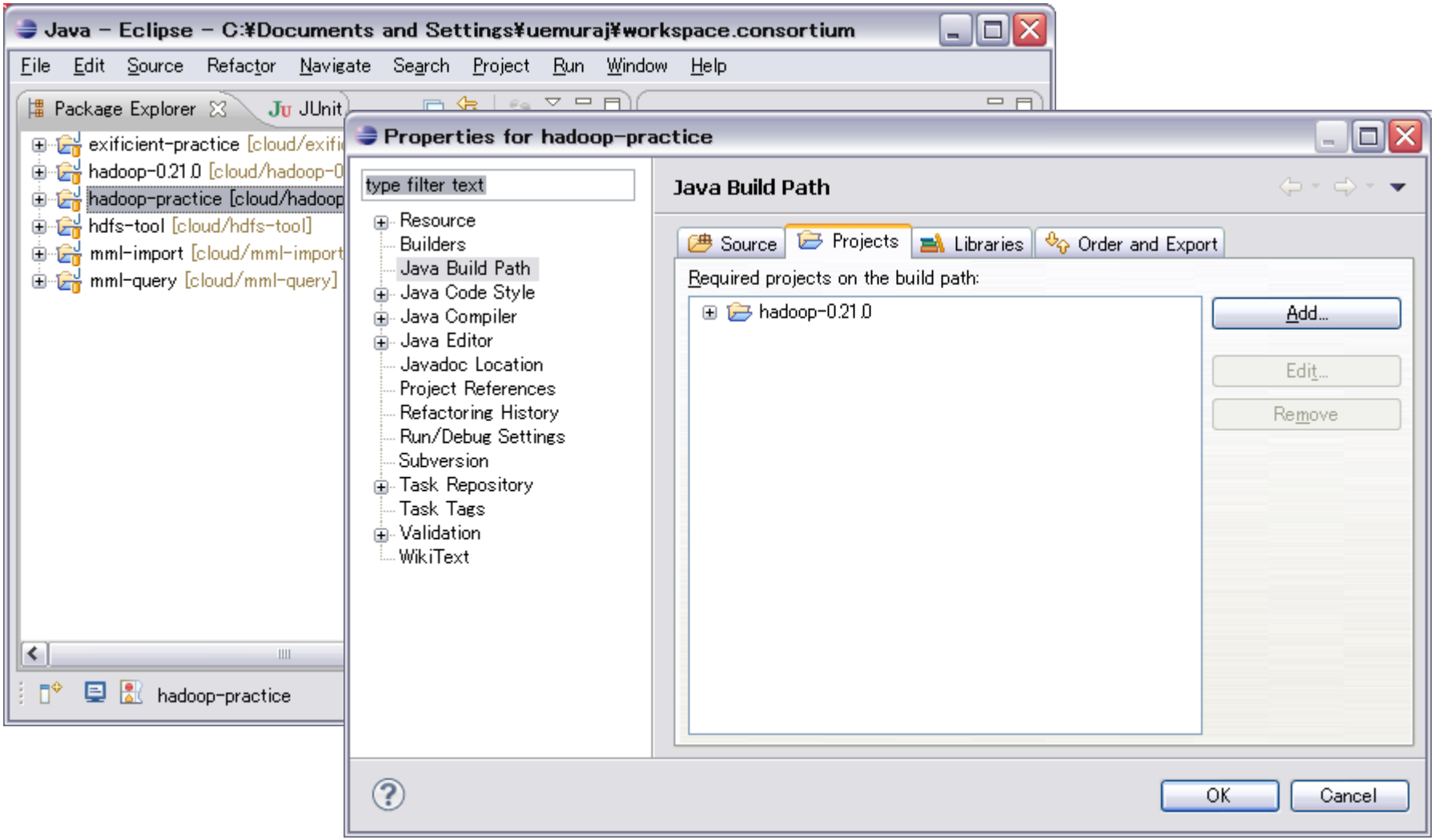
- プロジェクトからエクスポート



Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

開発環境の準備(６)

- プロジェクトを参照



Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

テストコードの記述

- Hadoopで記述した初めてのコード
  - Javaソースコードのimport文を数える
- テストされるコード – PackageCount.java
- テストするコード – PackageCountTest.java
- テスト結果 – PackageCountTest

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

圧縮データの入力(１)

- ZIP/TAR形式のデータを処理するInputFormatを実装
  - 入手したテスト用データをそのまま処理したい
  - MMLはXMLでありテキストデータ、まとめて圧縮すると圧縮率は高く、取り回しには便利

- 標準でもサポートされているが少し扱いが難しい
  - ZIPエントリで切り分けられたバイトデータがそのまま渡されるように見える

- それらしい名前の ZipFileRecordReader.java で検索してみる
  - 独自に実装されたものも見つかる(0.20系)
  - トレーニングとしてZIP/TARそれぞれ実装を試みる

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

圧縮データの入力(２)

- ZIP/TAR形式のデータを処理するInputFormatを実装
  - ZIPとTARどちらも非常に似た実装になる
  - GZIP圧縮されたTAR形式の場合の処理を紹介
- 実装したコード – TgzFileInputFormat.java
- テストするコード – TgzFileInputFormatTest.java
- テストするコード – XmlValidateMapper.java
- テスト結果 – TgzFileInputFormatTest

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

XMLデータの入出力(１)

- StreamXmlRecordReader を試してみる
  - mapred/contrib/streaming に用意されているのだが…
- テストするコード – StreamXmlRecordReaderTest.java
- 実行結果 – StreamXmlRecordReaderTest

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

XMLデータの入出力(２)

- org.w3c.dom.Documentを処理するSerializationを実装
  - Mapperの入力としてDOMオブジェクトが欲しい
  - Mapperを重ねれば可能
  - だがSerializationを実装するとノード間の通信にも使用される
  - XMLを良く圧縮できる実装があれば全体の効率UPが見込める
  - Writableを実装する方法よりシンプルに

- バイナリXMLを使用したSerializationを実装
  - XMLコンソーシアムに多数の公開資料
  - EXIficient – W3C Efficient XML Interchange (EXI) フォーマットのオープンソース実装を試す

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

XMLデータの入出力(３)

- org.w3c.dom.Documentを処理するSerializationを実装
  - WritableSerializationのソースコードを参考に

- 実装したコード – DocumentHolder.java
- 実装したコード – DocumentSerialization.java

- テストするコード – DocumentSerializationTest.java
- テストするコード – XmlDocumentMapper.java

- 実行結果 – DocumentSerializationTest

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

XMLデータの入出力(４)

- DocumentSerialization.java を使用して勉強会の課題を解く
  - 年齢と性別をキーに、その患者さんのBMI値を出力
- 実装したコード – BmiMapper.java
- テストするコード – BmiTest.java
- 実行結果 – BmiTest

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

まとめ

- まだまだ勉強すること、試すことが多くあります
  - 測定して実装方法による特性を調べたり
  - もっと他の手段がありそう、Hadoopを利用する他のソフトウェアも見てみたい
  - 実際にはそれらのソフトウェアを使うことになると思うが、このレベルの勉強は後々自分を助けると考える

- Hadoopをどのように稼動して利用するか
  - 「Amazon EC2」「GMOクラウド」「さくらVPS」での環境構築も試行錯誤中
  - 大規模な環境やセキュリティの話を聞くが、個人的には中小規模でも面白いことがありそうだと感じる

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.

さいごに

クラウド部会で一緒に勉強しましょう

Copyright © 2011 Advanced IT Consortium to Evaluate, Apply and Drive All Rights Reserved.