Using Pig as a data preparation language for large-scale mining software repositories studies: An experience report

By Weiyi Shang, Bram Adams, and Ahmed E. Hassan

Presentation by Wenhan Zhu (Cosmos)

Measure evolution of total number of LOC (Lines of Code) in different snapshots of a source control repository

- Extract information about line of code from repositories
- Transform the extrated information into LOC per snapshot
- Load the result in formats that can be used for analysis

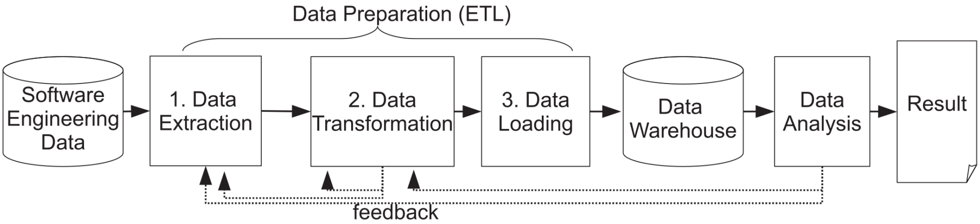

Fig. 1. ETL pipeline for large-scale MSR studies.

Figure taken from paper

Main contribution

Introducing Apache Pig as distributed platform for MSR studies

Pig

1. What is Pig?

   Pig is a high level Hadoop-based platform for analyzing large scale of data.

2. Why Pig?

   - Modularity
   - Scalability
   - Ease of deploying

Pig example code from paper

Sample code to measure evolution of LOC (Lines of Code) in a MSR study

```
RAWDATA = load ’ $inputdata ’ using ExtPigStorage ()
    as ( filename : chararray , filecontent : chararray ) ;
HISTORYLOG = foreach RAWDATA generate ExtractLog ( filename , filecontent ) ;
HISTORYVERSIONS = foreach HISTORYLOG generate ExtractVersions ( $0 ) ;
CODE = foreach HISTORYVERSIONS generate ExtractSourceCode ( $0 ) ;
LOC = foreach CODE generate GenLOC ( $0 ) ;
dump LOC ;
```

```
# Extracting Data
RAWDATA = load ’ $inputdata ’ using ExtPigStorage ()
    as ( filename : chararray , filecontent : chararray ) ;
# Transforming Data
HISTORYLOG = foreach RAWDATA generate ExtractLog ( filename , filecontent ) ;
HISTORYVERSIONS = foreach HISTORYLOG generate ExtractVersions ( $0 ) ;
CODE = foreach HISTORYVERSIONS generate ExtractSourceCode ( $0 ) ;
LOC = foreach CODE generate GenLOC ( $0 ) ;
# Loading Data
dump LOC ;
```

3 Case studies

- Study 1: Correlation between comment updates and bugs
  - is the change related to a bug?
  - does the change update source code comments?
- Study 2:Correlation between code clones and bugs
  - is the revision a bug fix?
  - is the method new or has it been deleted?
  - source code for very method
  - is the method cloned?
- Study 3: Evolution of the complexity of source code changes
  - number of changed LOC in Feature Introduction Modification (FI) changes

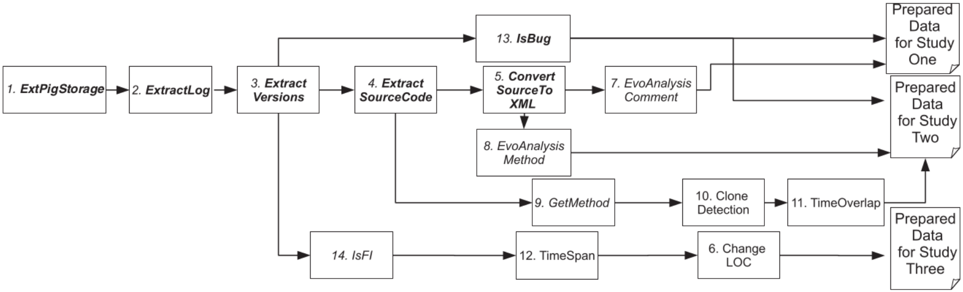

Fig. 7. Composition of the data preparation process for the three MSR studies performed with PIG. Modules with name in bold are used by more than one case study, whereas modules with name in italic are used by J-REX.

ExtPigStorage

```
CVSMETADATA = load ’ EclipseCvsData ’ using
ExtPigStorage () as ( filename : chararray , filecontent : chararray ) ;
```

ExtractLog

```
HISTORYLOG = foreach CVSMETADATA generate ExtractLog ( filename , filecontent ) ;
```

Extract Versions

```
HISTORYVERSIONS = foreach HISTORYLOG generate ExtractVersions ( $0 ) ;
```

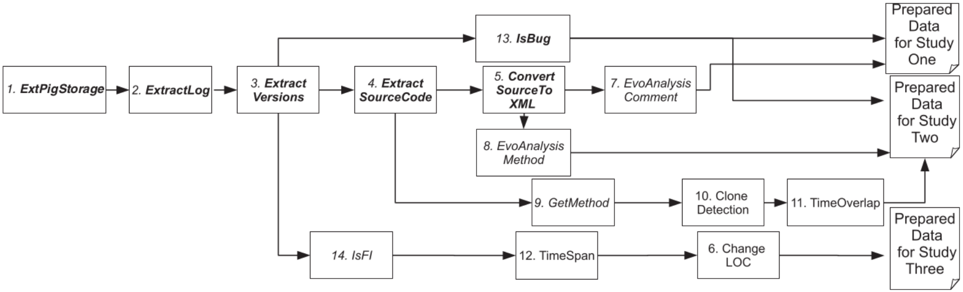

Fig. 7. Composition of the data preparation process for the three MSR studies performed with PIG. Modules with name in bold are used by more than one case study, whereas modules with name in italic are used by J-REX.

`IsBug`

```
BUGCHANGES = filter HISTORYVERSIONS by IsBug { $0 };
NOBUGCHANGES = filter HISTORYVERSIONS by not IsBug { $0 };
```

`Extract SourceCode`

```
CODE = foreach HISTORYVERSIONS generate ExtractSourceCode ( $0 ) ;
```

`Convert SourceToXML`

```
XMLS = foreach CODE generate ConvertSourceToX M L ( $0 ) ;
```

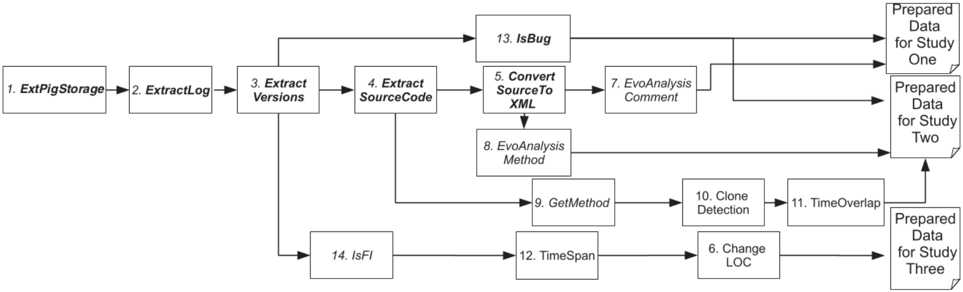

Fig. 7. Composition of the data preparation process for the three MSR studies performed with PIG. Modules with name in bold are used by more than one case study, whereas modules with name in italic are used by J-REX.

EvoAnalysis Comment

```
COMMENTEVO = foreach XMLS generate EvoAnalysisComment ( $0 ) ;
```

Prepared Data for Study One

```
BUGRESULT = join BUGCHANGES by $0 . $0 , COMMENTEV O by $0 . $0 ;
NOBUGRESULT = join NOBUGCHANGES by $0 . $0 , COMMENTEVO by $0 . $0 ;
dump BUGRESULT ;
dump NOBUGRESULT ;
```

Comparison to Hadoop

1. High level
2. Running Time
3. Migration Effort
4. Modularity

Things I like:

1. Introduces a high level platform that's easy to use for speeding up data preparation

2. Nicely explained diagrams for understanding the process of software engineering research

3. Detailed demo code for others to quickly pick up Pig as a platform

Things I would like to see improved:

1. Comparison with other ETL methods

2. Details of comparison methods between Hadoop and Pig

3. Typesetting on code segment

Discussions