**NIST Big Data Public Working Group (NBD-PWG)**

**NBD-PWD-2015/Use-case-#6b-data-mining-with-Weka.rr**

**Source: NBD-PWG**

**Status: Draft**

**Title: Big Data Use Case #6 Implementation, using NBDRA**

**Author: Afzal Godil (NIST), Wo Chang (NIST), Russell Reinsch (CFGIO), Shazri Shahrir**

To support Version 2 development, this Big Data use case (with publicly available datasets and analysis algorithms) has been drafted as a partial implementation scenario using the NIST Big Data Reference Architecture (NBDRA) as an underlying foundation. Feedback regarding the integration of the disruptive technologies described is welcomed.

## **Use Case #6b Data Mining**

### Introduction

# ​Big data is defined as high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. Both data mining and data warehousing are applied to big data and are business intelligence tools that are used to turn data into high value and useful information.

# The important differences between the two tools are the methods and processes each uses to achieve these goals. The data warehouse (DW) is a system used for reporting and data analysis. Data mining (also known as knowledge discovery) is the process of mining and analyzing massive sets of data and then extracting the meaning of the data. Data mining tools predict actions and future trends and allow businesses to make practical, knowledge-driven decisions. Modern data mining tools can answer questions that traditionally were too time consuming to be computed before.

### Given Dataset

2010 Census Data Products: United States (<http://www.census.gov/population/www/cen2010/glance/>)

Census data tools <http://www.census.gov/data/data-tools.html>

### Given Algorithms

1. Upon upload of the datasets to an HBase database, Hive and Pig could be used for reporting and data analysis.
2. Machine learning (ML) libraries in Hadoop and Spark could be used for data mining. The range of data mining tasks are: 1) Association rules and patterns, 2) Classification and prediction, 3) Regression, 3) Clustering, 4) Outlier Detection, 5) Time series analysis, 6) Statistical summarization, 7) Text mining, and 8) Data visualization

Note: all of these tasks are not required for every data mining implementation.

**Specific Questions to Answer:**

#1. What zip code has the highest population density increase in the last 5 years? And

#2. How is this correlated to unemployment rate in the area?

### Possible Development Tools Given for Initial Consideration

**Big-Data:**

Apache Hadoop, Apache Spark, Apache HBase, MongoDB, Hive, Pig, Apache Mahout, Apache Lucene/Solr, MLlib Machine Learning Library.

**Visualization:**

D3 Visualization, Tableau visualization.

**Languages:**

Java, Python, Scala, Javascript, JQuery.

**Use Case 6b: Data Mining with Weka**

Matched to the NIST BDRA M0437 PPT Activity and Functional Component View diagrams

Use Case Sections:

Primary actors and roles page 3

Level, preconditions and minimal guarantee page 3

Background on programming languages, tools in this use case page 3

General components of a data mining task page 3

**Main flow scenario, data mining with Weka,** extensions (dealing with errors) and alternate flows page 4

Subflows: visualization tools, model evaluation page 4

Licensing and TCO page 5

Legend:

Activities in this use case that match the BDRA M0437 Activities View diagram are noted in red type.

Functions in this use case that match the BDRA M0437 Functional Components View diagram are highlighted in yellow and noted in red type.

Primary roles and actors

1. Business Management Specialist. Should have previous experience with governance and compliance, and scientific method.
2. General Data Scientist. Should have experience with various types of data, NLP, ML, systems admin, DB admin, front end programming, optimization, data mgmt., data mining, and scientific method.
3. Developer / Engineer. Should have experience with product design, various data types, distributed data, systems admin, DB admin, cloud mgmt., back end programming, front end programming, math, algorithms, and data mgmt.
4. Research Scientist or Statistician. Readers should check Vol 2 for concurrency of additional roles. Scientist or statistician should have experience with machine learning, optimization, math, graphical models, algorithmss, Bayesian stats, data mining, statistical modeling, and scientific method.

Level, preconditions and minimal guarantee:

These use cases drafted in October 2015 represent the first steps in modeling of processes and as such still require iteration cycles to reach accuracy and completeness. This use case states a *generic analytical lifecycle*, we suggest that project managers should seek advice from professionals with backgrounds specific to the appropriate domain.

In an attempt to highlight essential dataflow possibilities and architecture constraints, as well as the explicit interactions between NBDRA key components relevant to this use case, this document makes some general assumptions. Certain functions and qualities will need to be validated in detail before the undertaking any proof of concept. In this use case, data governance is assumed. Security is not addressed.

Background on tools and languages in this use case

Weka: ML algorithms cover the range of mining tasks, especially good for time series. Differentiation may be its comprehensive pre-processing capabilities. Called from code (Java) or applied directly. GUI.

General components of a data mining project

1. Acquisition, collection and storage of data. Convert the data if necessary, providing a structure.
2. Organization of the data.
3. Identify features and decide on algorithms.
4. Prepare the data. Depending on which algorithm is determined to be most suitable, may take different approaches to preparing the raw data / type of workload.
5. Run the model
6. Evaluate model, validate results.

**Main Flow Scenario: Data Mining with Weka**

Acquire the data from the source. This activity matches the Big Data Application container, Collection oval in the BDRA Activities View.

The DB can be structured and formatted in MSFT Access. This activity matches the Platforms container, Index oval in the BDRA Activities View? This function matches the Infrastructure container, Storage box in the BDRAFCV.

Pre-processing: in this use case, the query task involves aggregates over certain metrics, vs. looking for individual events. Prior to mining; cleansing, transformation into appropriate format. In Weka, conversion to ARFF (ASCII text file) is suitable for processing. This activity matches the Big Data Application container, Preparation oval in the BDRA Activities View. This function matches the Big Data Application container, Transformation box in the BDRAFCV.

Integration: Hcatalog for metadata

Processing:

Processed attributes can (each) now be presented in a graphical representation (*descriptive* visualization activity), for the user to get a feel for the dataset. This activity matches the Big Data Application container, Visualization oval in the BDRA Activities View.

Run: Analysis visualizes the impact of population density on unemployment (/ second visualization activity). Two vintages of data, 2000 and 2010. The attributes are extracted and compared, and resulting values are plotted on 2 two-axis graphs. The first graph, for 2000 data has the unemployment variable on one axis and the density variable on the other axis. The second graph is plotted in the same way for 2010 data. This activity matches the Big Data Application container, Analytics oval in the BDRA Activities View. This function matches the Big Data Application container, Algorithms box in the BDRAFCV.

Concerns:

Possibility that Weka, R, or many of the advanced analysis technologies will not scale effectively for use cases with datasets large enough to require distributed systems. Alternate data mining technologies to consider: Conjecture, Mahout, MLBase, MLlib, Oryx, Samoa.

Subflow: Visualization

Matlab is a general purpose commercial visualization technology that enjoys current popularity in academic and government sectors as it is easy to learn to program. For scientific and numerical computing use cases involving multivariate and time data, Matlab performs very well in rand\_mat\_mul and pi\_sum benchmarks, nearly approaching C and Julia. Matlab slightly outperforms Python and R on rand\_mat\_stat, outperforms R on mandel and quicksort; but performs very poorly on fib and parse\_int benchmarks. Under some pressure to move toward open source systems, notably Python, the academic and government sectors are slowly moving away from commercial technologies.

Licensing and TCO

Weka: free under GNU.

Alternate flow scenario

RapidMiner would provide the needed level of mining capabilities with comparatively low licensing costs. This technology has a comparatively low requirement for analysts to have deep background in data mining or statistics.

Related references or reading

Census mining using Weka: <https://www.researchgate.net/publication/257882673_Census_Data_Mining_and_Data_Analysis_using_WEKA>

[*http://www.mastersindatascience.org/blog/open-source-tools-for-big-data-analysis/*](http://www.mastersindatascience.org/blog/open-source-tools-for-big-data-analysis/)

Census analysis: <http://www.iosrjournals.org/iosr-jce/papers/Vol14-issue2/A01420105.pdf?id=6815>

Change in median income: <http://www.census.gov/did/www/saipe/data/highlights/files/2013/F2_MP_13.pdf>

Density change pdf: <http://www2.census.gov/geo/pdfs/maps-data/maps/thematic/density.pdf>