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

**NBD-PWD-2015/M0444 v2**

**6b,Data.mining.abbreviated.rr.ss**

**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 analytic algorithms) has been drafted as a partial implementation scenario using the NIST Big Data Reference Architecture (NBDRA) as an underlying foundation. 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. Feedback regarding the integration of the disruptive technologies described is welcomed.

## **Use Case #6: a) Data Warehousing and b) 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. 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/>)

### 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.

**Analysis of use case 6b Data Mining, against the NIST RA perspective, namely the M0437 PPT Activity and Functional Component View diagrams:**

Sections:

Background on programming languages page 3

Background on data mining tools. Two libraries are most suitable… page 3

The analysis task using Weka page 6

Background on visualization tools page

TCO page

Background on Programming Languages

Python, like R, is a free and open source scripting language that can be used with MapReduce for analysis. Python has been steadily gaining share of business applications though R continues to maintain a larger base in analytical processing ([Piatetsky 2013](#_ENREF_40)). As the statistical and visualization toolboxes in Python have steadily improved over recent years, academics and government agencies are finding it to be a viable option for model building and structured data operations. Python is slightly easier to learn than R, but does not have quite the depth in development community as R (Jain 2014). *The following are examples of python packages that can be considered foundations in data science;*

1. *Numpy , numerical packages*
2. *Scipy, scientific computing packages*
3. *Matplotlib, presentation packages*

*Other data science foundations include distance metrics, dimension reduction, and* ***transformation.***

The R Project is both a statistical language suite and a development environment that is a cost effective alternative to commercial analysis software as it is open source. Arguably the most powerful analytic environment available, R was used by NIST to calculate the oil flow rate in the Gulf of Mexico during the BP oil rig disaster. R is used heavily by academics; but has weaknesses in exploration, discovery and visualization functionality. R is a low level language making it more difficult to learn than Python; for use cases requiring customized data analysis and mining tasks, R is now seeing competition from Python.

Background on Data Mining Tools

E View, Excel, Minitab, SAS + Eminer, SPSS Modeler, Stata, Statistica

Knime Big Data Extension and Cluster Execution: large standalone framework good for time series. Integrates with R and Weka. OS desktop version. Not a highly rated architecture but one of the top five in product strength. Knime has a history of use in government as well as the life sciences, communications and education industries. KNIME has a strong support community and a working relationship with Actian (a business intelligence vendor). KNIME is good for users looking for capabilities to build filtering, data access, and advanced analytics solutions from scratch; but not simulations. KNIME is a top advanced analytics product with an extensive range of functions yet it remains user friendly.

Mahout. Strength in clustering and classification.

MLlib:

Prognoz: time series. Headquartered in Washington DC?

Radoop: a commercial implementation of RapidMiner.

RapidMiner: Wide range of capabilities, written in Java, can be integrated with almost any major operating system or platform. Can integrate with other products or operate standalone. Pulls ML from Weka. Integrate with R for statistical modeling or use its native Rapid-I scripting library. Also one of the top five in algorithm strength. Similar to KNIME but weaker in platform management. RapidMiner is known to be easy to use; and many customers employ the software for text analysis. As its name suggests, RapidMiner is fast, boasting an ability to solve scenarios in five minutes. The product has been designed primarily for business users. Headquartered in Burlington, MA.

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

Wise.io: time series

Zaitun: time series.

Dryad and Spark are lowest level processing layer technologies that extend Hadoop, generalizing and potentially eliminating the need for MapReduce, depending on what type of work a user needs to perform. Both Dryad and Spark use graph technology and follow many principles of MPP. Developed by MSFT, Dryad powers Bing; and has strong capabilities for chaining operators. Capable of efficiently expressing several types of computations, Dryad solves one of the problems in MapReduce, that of insufficient data sharing in between Jobs.

Generally speaking, programmers consider Dryad and the other low layer technologies such as Spark and Hadoop as too painful to use; in other words, very complicated and cumbersome, and this restricts their adoption. Dryad LINQ provides a more familiar language for programmers to write code.

Spark: is a newer OS Scala compatible analytics platform / DAG that extends Dryad. More simply, it is a computation engine,

In contrast to MapReduce’s batch architecture that is appropriate for heavy / large processing tasks, Spark’s multipurpose engine can process writes at the speed of memory, making it a better option than MapReduce for quick, lighter queries. Spark is one of the in memory technologies providing a high level SQL like interface for interacting with HDFS. Spark is very popular general purpose engine for data reaching into the hundreds of TB. Still immature / has no security. Hive tez vs. spark sql vs. hive on spark. (screenshot.) vs. impala.

Analytic Engines with Hadoop Interoperability

Concurrent: fills a unique area. App dev tool, new money, check Pepperdata.

Datameer is a native Hadoop application that provides a front end that shelters users from some of the complex programming required in MapReduce and Hadoop. Using a graphical interface and wizard driven technology, Datameer offers a highly desired experience commonly referred to as “self service.” Datameer uses a spreadsheet approach but does not integrate natively with Microsoft Excel. API interoperability connections are made through ODBC, JDBC and a metadata catalogue. As JDBC, like SQL, is designed for interfacing data stored in row formats, users face the same issue found in old fashioned BI technologies, and will need to design their entire output artifacts before query; though the company is working on new tools for processing the dirty types of data found in OS use cases.

An average contract with Datameer is $100,000. With a stable of customers that includes Citi, Sears, Visa and Workday, Datameer has impressive top line growth and can be considered a low risk to organizations looking for an analytic platform that is not as big as the established vendors SAS, EMC and Teradata.

Karmasphere is classified as an analytic pure play, meaning the company specializes in one thing and doing it well. Similar to Datameer and Platfora in this respect, and in the respect that the three companies each have some capability for visualization of data, Karmasphere Studio differentiates itself as an integrated development environment (IDE) for MapReduce Job development and deployment. Querying of large data sets can be accomplished through Hive, Pig, or with R. The Karmasphere Analyst solution provides visualization functionality for Microsoft Excel and Tableau.

Professional product editions are available for annual subscription and a community edition is free. An alternative, on demand pricing option is available for Amazon EMR integrations. The company is product oriented and provides some in demand type functionality such as visual mechanisms for code stringing; but as a startup, Karmasphere does not appear to be on a trajectory to experience above average growth.

-----------------------------------------------------------------------------------------------------------------

Mining with Weka: Exploration; Model

Prior to Mining: cleansing, transformation into appropriate format.

Pre-processing: DB could be structured and formatted in MSFT Access. Convert to ARFF (ASCII text file) for processing. Processed attributes can (each) now be presented in a graphical representation (/ first visualization activity).

Suggest to state a *generic analytical lifecycle*, probably good to get advise from somebody from Public Health background. Like here I agree that we would like to do some *descriptive/visualization* to get a feel of the data first. Then some *Insights*, to find interesting variables or contributors, then probably *Prediction*, after which *Prescription* of how to solve the problem , after all that we need to do *Impact Analysis* from action which has been taken.

Attempt to visualize the impact of population density on unemployment (/ second visualization activity). Two vintages, 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.

-----------------------------------------------------------------------------------------------------------------

Vis12: a number of visualization techs work on time data. cubism, envision, rickshaw, all do time series analysis; d3, datahero, exhibit, gnuplot, js infovis, keshif, matlab, matplotlib, mstr, miso project, piety, pivotviewer, qliksense, rickshaw, raw, timeline.js, quadrigram, crossfilter. Quadrigram and miso project are “core” vis12n technologies.

Matlab is a commercial vis12n technology that has enjoyed recent popularity in academic and government sectors as it is easy to learn to program. For scientific and numerical computing use cases, 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…

TCO

Rapidminer would provide the needed level of mining capabilities with comparatively low licensing costs. Comparatively low requirement for the analyst to have deep background in data mining or statistics.

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

Correlogram

Tigerweb

Background on integration technologies. Hcatalog for metadata

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>