International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-3, March 2016
ISSN: 2395-3470
www.ijseas.com

# DATA MINING TOOLS –APPLICABLE ON STORAGE SYSTEMS

**Nagamani Mutteni**

**Asst.Prof,MERI,New Delhi**

**Abstract:** The expeditious development of information technology and adaptability of service technologies created a regime in various fields appreciably. The growing interest in usage of data for analysis has brought an indispensable enhancement in data mining field. Data mining and it's applications can contemplate as one of the transpiring and promising technological developments that provide efficient means to access various types of data and information available globally. These applications also aid in decision making. This paper discusses about technical specifications and features of some of the open source data mining tools.

**Keywords: Data, Data Mining, Data Mining Tools, Apache Mahout, Ratter, Criptella**

**Introduction:**

There has been a substantial increase in amount of information and data which is stored in electronic format since last few decades. The size of data base has been in the process of continuous increment and has reached up to 3.77 zettabytes in 2016 according to IDC. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world doubles every 12 to 18 months. As the information grows on, the retrieval of information gets complicated but data mining can substitute. Data mining refers to extracting or mining knowledge from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. This can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

**Data Mining Tools:** Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences.

Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult.Some of the open source data mining tools available are

1) Apache Mahout
2) Scriptella
3) Rattle
4) NLTK
5) Weka
6) jHepWork

1) **Apache Mahout** is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. Many of the implementations use the Apache Hadoop platform. Mahout also provides Java libraries for common maths operations (focused on linear algebra and statistics) and primitive Java collections. A mahout is one who drives an elephant as its master. The name comes from its close association with Apache Hadoop which uses an elephant as its logo.

**Hadoop** is an open-source framework from Apache that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.

Apache **Mahout** implements popular machine learning techniques such as:

1. Recommendation
2. Classification
3. Clustering

Apache Mahout started as a sub-project of Apache's Lucene in 2008. In 2010, Mahout became a top level project of Apache

**Features of Mahout**

The primitive features of Apache Mahout are listed below.

1. The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment. Mahout uses the Apache Hadoop library to scale effectively in the cloud.
2. Mahout offers the coder a ready-to-use framework for doing data mining tasks on large volumes of data.
3. Mahout lets applications to analyze large sets of data effectively and in quick time.
4. Includes several MapReduce enabled clustering implementations such as k-means, fuzzy k-means, Canopy, Dirichlet, and Mean-Shift.
5. Supports Distributed Naive Bayes and Complementary Naive Bayes classification implementations.
6. Comes with distributed fitness function capabilities for evolutionary programming.
7. Includes matrix and vector libraries.

**Applications of Mahout**

1. Companies such as Adobe, Face book, LinkedIn, Foursquare, Twitter, and Yahoo use Mahout internally.
2. Foursquare helps you in finding out places, food, and entertainment available in a particular area. It uses the recommender engine of Mahout.
3. Twitter uses Mahout for user interest modeling.
4. Yahoo! uses Mahout for pattern mining.

2) **Scriptella** is an open source ETL (Extract-Transform-Load) and script execution tool written in Java. Its primary focus is simplicity. It doesn't require the user to learn another complex XML-based language to use it, but allows the use of SQL or another scripting language suitable for the data source to perform required transformations. Scriptella does not offer any graphical user interface.

**Features**

1. Support for **multiple datasources** (or multiple connections to a single database) in an ETL file.
2. Support for many useful **JDBC features**, e.g. parameters in SQL including file blobs and JDBC escaping.
3. Low memory usage is one of the primary goals.
4. Support for **evaluated expressions and properties** (JEXL syntax)
5. Support for **cross-database ETL scripts** by using <dialect> elements
6. **Transactional execution**
7. **Error handling** via <on error> elements
8. **Conditional scripts/queries execution** (similar to Ant if/unless attributes but more powerful)
9. **Easy-to-Use** as a standalone tool or Ant task. No deployment/installation required.
10. **Easy-To-Run** ETL files directly from Java code.

**Applications:**

1. JDBC/ODBC compliant driver.
2. Service Provider Interface (SPI) for interoperability with non-JDBC DataSources and integration with scripting languages. Out of the box support for JSR 223 (Scripting for the Java Platform) compatible languages.
3. CSV, TEXT, XML, LDAP, Lucene, Velocity, JEXL and Janine providers. Integration with Java EE, Spring Framework, JMX and JNDI for enterprise ready scripts.

**3**. **Rattle GUI** is a free and open source software ( GNU GPL v2) package providing a graphical user interface (GUI) for data mining using the R statistical programming language. Rattle can be used for statistical analysis, or model generation. Rattle allows for the dataset to be partitioned into training, validation and testing. The dataset can be viewed and edited. There is also an option for scoring an external data file.

**Features**

1. File Inputs = CSV, TXT, Excel, ARFF, ODBC, R Dataset, RData File, Library Packages Datasets, Corpus, and Scripts.
2. Statistics = Min, Max, Quartiles, Mean, St Dev, Missing, Medium, Sum, Variance, Skewness, Kurtosis, chi square.
3. Statistical tests = Correlation, Wilcoxon-Smirnov, Wilcoxon Rank Sum, T-Test, F-Test, and Wilcoxon Signed Rank.
4. Clustering = KMeans, Clara, Hierarchical, and BiCluster.
5. Modeling = Decision Trees, Random Forests, ADA Boost, Support Vector Machine, Logistic Regression, and Neural Net.
6. Evaluation = Confusion Matrix, Risk Charts, Cost Curve, Hand, Lift, ROC, Precision, Sensitivity.
7. Charts = Box Plot, Histogram, Correlations, Dendrograms, Cumulative, Principal Components, Benford, Bar Plot, Dot Plot, and Mosaic.
8. Transformations = Rescale (Recenter, Scale 0-1, Median/MAD, Natural Log, and Matrix) - Impute ( Zero/Missing, Mean, Median, Mode & Constant), Recode (Binning, Kmeans, Equal Widths, Indicator, Join Categories) - Cleanup (Delete Ignored, Delete Selected, Delete Missing, Delete Obs with Missing)

Rattle also uses two external graphical investigation / plotting tools. Latticist and GGobi are independent applications which provide highly dynamic and interactive graphic data visualisation for exploratory data analysis.

**Applications**

1. It finds its use in statistical applications like Skewness, Kurtosis,etc.
2. It can be used in business applications to determine Cost Curves, Risk Charts, etc
3. It can be used to pictorially represent data.

**4. NLTK**

**Natural Language Tool Kit** is a Python package that implements many standard NLP data structures and algorithms. It was first developed in 2001 as part of a CL course at University of Pennsylvania and led by Steven Bird, Edward Loper, Ewan Klein. It is one of the open-source data mining tools.

**Features**

1. Accessing corpora - Standardized interfaces to corpora and lexicons
2. String processing - Sentence and word tokenizers
3. Stemmers
4. Part-of-speech tagging - Various part-of-speech taggers
5. Classification - Decision tree, maximum entropy
6. K-means
7. Chunking - Regular expressions, named entity tagging

**Applications**

1. It can be used for string processing.
2. It can be used to form clusters.
3. It can be used in information retrieval through data mining algorithms.

**5. Weka**

**Waikato Environment for Knowledge Analysis** Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

**Features**

1. It provides many different algorithms for data mining and machine learning
2. It is open source and freely available
3. It is platform-independent
4. It is easily useable by people who are not data mining specialists
5. It provides flexible facilities for scripting experiments
6. It has kept up-to-date, with new algorithms being added as they appear in the research literature.

**Applications**

1. Weka is best suited for mining association rules .
2. It is stronger in machine learning techniques.
3. It is suited for machine Learning.
4. It is also suitable for developing new machine learning schemes.

5. Weka loads data file in formats of ARFF, CSV, C4.5, binary. Though it is open source, Free, Extensible, can be integrated into other java packages.

**6.jHepWork**

jHepWork is a full-featured object-oriented data analysis framework for scientists that takes advantage of the Jython language and Java. Jython macros are used for data manipulation, data visualization (plotting 1D and 2D histograms), statistical analysis, fits, etc. Data structures and data manipulation methods integrated with Java and JAIDA FreeHEP libraries combine remarkable power with a very clear syntax. jHepWork Java libraries can also be used to develop programs using the standard JAVA, without Jython macros. jHepWork consists of two major libraries: jeHEP (jHepWork IDE) and jHPlot (jHepWork data-analysis library). Both are licensed by the GNU General Public License (GPL).

**Features**

1. Programs written using jHepWork are usually short due the simple Python syntax and highlevel constructs implemented in the core jHepWork libraries.
2. jHepWork is an open source product which is implemented 100 percent in Java. Since it is fully multiplatform, it does not require installation and can be run on any platform where Java is installed.
3. jHepWork is seamlessly integrated with Java-based Linear Collider Detector (LCD) software concept and it has the core based using FreeHEP libraries and other GNU-licensed packages.
4. It offers a full-featured, extensible multiplatform IDE implemented in Java.

**Applications**

1. As a front-end data-analysis environment, jHepWork helps to concentrate on interactive experimentation, debugging, rapid script development and finally on workflow of scientific tasks, rather than on low-level programming.
2. It can be used to develop a range of data-analysis applications focusing on analysis of complicated datasets, histograms, statistical analysis of data, fitting.
3. jHepWork Java libraries can also be used to develop programs using the standard JAVA, without Jython macros.

**Conclusion**

Open-source data mining suites of today have evolved to a great extent. They provide flexibility either through visual programming within graphical user interfaces or prototyping by way of scripting languages. They offer nice graphical interfaces, focus on the usability and interactivity, support extensibility through augmentation of the source code or through the use of interfaces. The study presented description of various open source data mining tools enlisting their features and applications. This paper provides an insight in future which, instead of supporting a specific area, the use of these tools can be extended to more fields with more technical improvements.

**References**

1. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.
2. Witten, I.H., Frank, E.: “Data Mining: Practical machine Learning tools and techniques”, 2nd addition,Morgan Kaufmann, San Francisco(2005).
3. Comparative Study of Data Mining Tools Kalpana Rangra Dr. K. L. Bansal.
4. http://www.r-project.org/
5. krish@cs.toronto.edu / t4peruma@cdf.toronto.edu