International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 04 || April 2017 || PP. 60-65

Comparative Analysis of Open source Business Intelligence tools for Crime Data Analytics

Emmanuel Ahishakiye[1], Elisha Opiyo Omulo[2], Danison Taremwa[3],Ruth Wario[4]
[1, 2]*(School of Computing and Informatics, University of Nairobi, P.O. Box 30197 – 00100, GPO Nairobi-Kenya)*
[3]*(Department of Computer science, Kyambogo University, P.O. Box 1, Kyambogo, Kampala – Uganda)*
[4]*(Department of Computer Science and Informatics, Faculty of Natural and Agricultural Sciences, University of Free State, Private Bag X13, Kestell 9866, Republic of South Africa)*

Abstract: Law enforcement agencies like other organizations are facing a difficult task of handling and making use of crime data which is in different formats that is generated every day which would otherwise help them in effective crime management. Developing a low cost Business Intelligence system for crime data analytics requires low cost development tools and this is where open source business intelligence tools come to rescue. Therefore there is a need to identify an efficient and effective open source business intelligence tool for the implementation of a Business Intelligence System for crime data analytics. This paper discussed five open source BI tools; Apache Hadoop, Jaspersoft, Pentaho, SpagoBI and vanilla. From the analysis, Apache Hadoop is recommended by this research for crime data analytics because it has some functionalities which are not found to other open source tools which includes but not limited to distributed storage and processing of big data sets, Ability to store and process huge amounts of any kind of data quickly, fast computing power, Fault tolerance, Flexibility (processes structured, semi structured and un structured data) and Scalability. Pentaho and SpagoBI come the second with very good number of features, then vanilla follows and lastly Jaspersoft. The analysis of the features was only limited to open source BI tools and therefore the commercial versions of the tools were not considered in this study.

Keywords: Crime data analytics, Law Enforcement Agencies, open source Business Intelligence Tools

I. Introduction

Technology now allows us to capture and store vast quantities of data. Finding patterns, trends, and anomalies in these datasets, and summarizing them with simple quantitative models, is one of the grand challenges of the information age - turning data into information and turning information into knowledge. As the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly. This kind of large data also called as Big Data and 80% of the world's data is now in unstructured formats, which is created and held on the web. Over the next decade there will be 45 times more data than today [1]. Lying hidden in all this data is information, potentially useful information that is rarely made explicit or taken advantage of. Intelligently analyzed data is a valuable resource. It can lead to new insights and, in commercial settings, to competitive advantages.

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences, crime trends and other useful information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages, reduced crime activities and public safety and other benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs [2].

Business Intelligence (BI) tools extract and interpret from the mass of business data that your organisation collects and provide you with information that can be useful to you. Open source business intelligence and reporting tools provide a rich feature set ready for enterprise use. End users should do a thorough comparison and select the tool that best meets their needs. Some of the tools distinguish themselves by specific features such as integration with machine learning, or availability of virtual machine and cloud images.

I. Literature Survey

A. Data analytics

Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses. Data analytics can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals - all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources [3]. Crime data analysis is used to analyze, map and visualize crime incidents or crime pattern to have an idea for predicting the crime occurrence. Crime data analytics thus helps the security as well as police to accommodate their resources accordingly for preventing crime [4].

B. Open Source Software

Open-source software (OSS) is computer software with its source code made available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. [5]Open-source software may be developed in a collaborative public manner. According to scientists who studied it, open-source software is a prominent example of open collaboration. Open source software is usually easier to obtain than proprietary software, often resulting in increased usage. Additionally, the availability of an open source implementation of a standard can increase adoption of that standard. [6] It has also helped to build developer loyalty as developers feel empowered and have a sense of ownership of the end product. [7] Moreover, lower costs of marketing and logistical services are needed for OSS. OSS also helps companies keep abreast of technology developments. It is a good tool to promote a company's image, including its commercial products. [8] The OSS development approach has helped produce reliable, high quality software quickly and inexpensively.

C. Open Source versus Commercial Software Tools

Many organizations are looking to reduce costs in their large Business Intelligence (BI) deployments and they are hoping that open source BI gives them greater leverage for their money. The most relevant benefit of using open source tools is that they are free and allow access to the source code, with the possible modification of the various modules. One of the advantages of open source platforms are that if they do not serve the needs of an organization, then they can be replaced by other platforms without a cost; but this does not happen with commercial platforms [9]. The open source tools generally require lower system requirements than commercial applications and this is justified in assuming that those who cannot invest in software may not invest in hardware. it is believed that Open Source BI platforms will evolve much faster than commercial ones since they are not constrained by compatibility problems and rigid (or even obsolete) architectures. More on that, Open Source solutions can exploit the contributions of the Open Source development community that relies on hundreds of programmers and designers as well as on the direct involvement of researchers [10]. One of the biggest problems with commercial solutions is that all the costs are born upfront by the customer before there is any reward. Other demerits of commercial BI platforms are: high acquisition costs, the requirement to be connected to the sellers, typically require more powerful hardware and the difficulty of transition to other platforms, taking into account the costs and terms of the contract [11].

D. Selected Open Source Business Intelligence Tools

1. Apache Hadoop

Apache Hadoop is an open-source software framework used for distributed storage and processing of big data sets using the MapReduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that

hardware failures are common occurrences and should be automatically handled by the framework.[12] The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, [13] where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.[14]

The base Apache Hadoop framework is composed of the following modules:

- Hadoop Common - contains libraries and utilities needed by other Hadoop modules;
- Hadoop Distributed File System (HDFS) - a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
- Hadoop YARN - a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications; and
- Hadoop MapReduce - an implementation of the MapReduce programming model for large scale data processing.

The term Hadoop has come to refer not just to the base modules above, but also to the ecosystem,[15] or collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache ZooKeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache Oozie, Apache Storm. The basic principle of working behind Apache Hadoop is to break up unstructured data and distribute it into many parts for concurrent data analysis. Big data applications using Apache Hadoop continue to run even if any of the individual cluster or server fails owing to the robust and stable nature of Hadoop. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. With big data being used extensively to leverage analytics for gaining meaningful insights, Apache Hadoop is the solution for processing big data. Apache Hadoop architecture consists of various Hadoop components and an amalgamation of different technologies that provides immense capabilities in solving complex business problems.

2. JasperReport

JasperReport is one of the most popular and widely used open source reporting tools. It is used in hundreds of thousands production environments, and features both community and commercially-supported versions. JasperReport consists of several components including the JasperReport Library, JasperReport Studio, and JasperReport Server. The library includes the entire core Java classes and APIs powering JasperReport. The ETL, OLAP, and server components provide JasperReport with important tools for enterprise environments, making it easier to integrate with the existing IT architecture of organizations. JasperReport is supported by excellent documentation, a wiki, and additional resources. Written in Java, JasperReport runs on Windows, Linux, and Mac is licensed under AGPL [16].

3. Pentaho

Pentaho is a complete business intelligence suite, covering a gamut of use cases from reporting to data mining. The Pentaho BI suite encompasses several open source projects, of which Pentaho Reporting is one of them. Like the other tools, Pentaho Reporting has a rich feature set, ready for use in enterprise organizations. The Pentaho BI suite also contains the Pentaho BI Server. This is a J2EE application which provides an infrastructure to run and view reports through a web-based user interface. Pentaho is supported through many community resources such as documentation, wiki, and more. The tool runs on Java Enterprise Edition and can be used on Windows, Linux, and Mac [17].

4. SpagoBI

SpagoBI is an Open Source Business Intelligence suite, belonging to the free/open source SpagoWorld initiative, founded and supported by Engineering Group. It offers a large range of analytical functions, a highly functional semantic layer often absent in other open source platforms and projects, and a respectable set of advanced data visualization features including geospatial analytics. SpagoBI is released under the Mozilla Public License, allowing its commercial use. SpagoBI is hosted on OW2 Forge managed by OW2 Consortium, an independent open-source software community. SpagoBI Tools include for example: reporting, charts, cockpits, data-mining, ETL, and many more. SpagoBI can integrate with many other tools, such as KeyRock identity manager, Orion Context Broker, and CKAN, the popular and widely used open data portal. It is certified for environments including Wildfly 8, 10 and JBoss EAP 7 [18].

5. Vanilla

Vanilla is an open source business intelligence platform developed in PHP language and is available in a single version, community. Vanilla is a complete platform since it integrates all main functionalities of business intelligence. The platform allows to create reports, perform analysis, generate and analyze data tables, charting several, dashboards, ad-hoc queries, integrate OLAP and ETL processes, define KPIs, data export and use of procedures of Data Mining. BIRT and iReports are used for the purpose of metadata integration. Vanilla allows filing several separate projects, and an historical book. The platform is available in a version for mobile operating system Android. Vanilla is available for Linux, Windows and UNIX operating systems. Vanilla was the first open source tool to provide a mobile version, in version 3.4, where is possible to browse reports and run dynamic reports from smart phones. It is also possible to browse OLAP cubes [19].

II. Methodology

The objective of this study is to find the most effective and efficient open source business intelligence tool for crime data analytics. This will be done by analyzing the features of each of the tools and then comparing the features of the tools with the needs of the ever increasing data (Big Data). The features that are considered are; reports, graphics, dashboards, OLAP, ETL, Data mining, KPIs, data export, GEO/GIS, adhoc reporting, ability to run on multiple operating systems (Linux, windows, Unix, Mac), support of java programming, distributed storage and processing and data recovery in case of cluster hardware failure.

A. Sources of the selected open source BI Tools

All the BI tools considered in this study are open source and therefore available free of charge for download. Apache Hadoop, Jaspersoft, Pentaho, SpagoBI and Vanilla can be downloaded from [12], [16], [17], [18] and [19] respectively.

B. Criteria used for evaluation

The criteria used for the comparative study of the selected BI tools are the features they have that enable them to perform the business intelligence tasks. Each of the open source BI tool was studied to determine the most tool for crime data analytics. This criteria used have been used by other researchers [10] [20] [21] [22] and the features considered took into account the available crime data which is in different formats (structured, semi structured and unstructured crime data). The following are the Business Intelligence Indicators/ Features that were considered in this study.

- Reports
- Dashboards
- OLAP – Online Analytical Processing
- ETL – Extraction, transformation and loading
- Data Mining
- KPIs – Key Performance Indicators
- GEO/ GIS – Geo Information System
- Ad-Hoc Queries

- Linux
- Windows
- Unix
- Mac
- Java
- Distributed storage & Processing
- Fault tolerance
- Scalability

III. Evaluation of Open Source Business Intelligence Tools

The evaluation of the BI tools is based on the features as shown in table 1 below.

Table 1: Open Source Business Intelligence Platforms Features

Features	Apache Hadoop	Jaspersoft	Pentaho	SpagoBI	Vanilla
Reports	✓	✓	✓	✓	✓
Graphics	✓	✓	✓	✓	✓
Dashboards	✓	✓	✓	✓	✓
OLAP	✓	✓	✓	✓	✓
ETL	✓	✓	✓	✓	✓
Data mining	✓	x	✓	✓	✓
KPIs	✓	x	✓	✓	✓
Data export	✓	✓	✓	✓	✓
GEO/GIS	✓	✓	✓	✓	x
Adhoc queries	✓	✓	✓	✓	✓
Linux	✓	✓	✓	✓	✓
Windows	✓	✓	✓	✓	✓
Unix	✓	x	✓	✓	✓
Mac	✓	✓	✓	✓	✓
Java	✓	✓	✓	✓	x
Distributed storage & Processing	✓	x	x	x	x
Fault tolerance	✓	x	x	x	x
Scalability	✓	x	x	x	x

From table 1 above, Apache Hadoop has most features compared to Jaspersoft, Pentaho, SpagoBI and Vanilla. It also allows distributed storage and processing of big data sets, Ability to store and process huge amounts of any kind of data quickly, fast computing power, Fault tolerance, Flexibility (processes structured, semi structured and un structured data) and Scalability. Pentaho and SpagoBI come the second with very good number of features, then vanilla follows and lastly Jaspersoft. The analysis of the features was only limited to open source BI tools and therefore the commercial versions of the tools were not considered in this study.

IV. How This Study Will Help Big Data System Developers And Consumers

This study will act as a guide to both the big data system developers and consumers as it will act as a road map in selection of the best and most efficient open source BI tool for crime data analytics. The selection of the tool to use will depend on the problem at hand to be solved but the researchers believe that this study will help the developers and consumers in making their decisions on which open source BI tool to use in their data analytics.

V. Conclusion and Future Work

The researchers concluded that the Open Source Business Intelligence platforms are growing both in features, quality and visual appeal. Of the six tools reviewed, Apache Hadoop is recommended by this research for crime data analytics because it has some functionalities which are not found to other open source tools which includes but not limited to distributed storage and processing of big data sets, Ability to store and process huge amounts of any kind of data quickly, fast computing power, Fault tolerance, Flexibility (processes structured, semi structured and un structured data) and Scalability which are key to handle the ever increasing data which is in different forms (structured, semi structured and unstructured data). Finally, Hadoop is simple and provides a fast developing environment Its Map Reduce scheme keeps the flowchart simple for programmers and its installation can be easily modified in little time .in case of utility or a user does not have the resources to install the platform, it is easily accessible on the cloud too. Hadoop represent an increasingly important approach for data-intensive computing.

The future work of the researchers would involve new studies and implementations of BI with Data warehousing to create a technological tool to support the decision-making for law enforcement agencies in crime management using Apache Hadoop framework by taking this paper as a base.

VI. References

[1]. Dr. Rakesh Rathi and Sandhya Lohiya . Big Data and Hadoop. International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014). Vol. 2, Issue 2, Ver. 2 (April - June 2014).
[2]. Lisa Martinek and Craig Stedman (2013). Big data analytics http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
[3]. Craig Stedman. Data analytics (DA). http://searchdatamanagement.techtarget.com/definition/data-analytics
[4]. Vikas Grover, Richard Adderley, Max Bramer, Review of Current Crime Prediction Techniques
[5]. St. Laurent, Andrew M. (2008). Understanding Open Source and Free Software Licensing. O'Reilly Media. p. 4. ISBN 9780596553951.
[6]. US Department of Defense (2016). "Open Source Software FAQ". Chief Information Officer. Retrieved 22 July 2016.
[7]. Sharma, Srinarayan; Vijayan Sugumaran; Balaji Rajagopalan (2002). "A framework for creating hybrid-open source software communities" (PDF). Info Systems Journal. 12: 7–25. doi:10.1046/j.1365-2575.2002.00116.x.
[8]. Landry, John; Rajiv Gupta (September 2000). "Profiting from Open Source". Harvard Business Review. doi:10.1225/F00503 (inactive 2017-01-18).
[9]. Madureira, L. (2012). Computational Intelligence and Decision Making: Trends and Applications.
[10]. Matteo Golfarelli. Open Source BI Platforms: a Functional and Architectural Comparison. DEIS, University of Bologna, Viale Risorgimento 2, Bologna, Italy.
[11]. Bernardino, J., & Tereso, M. (2013). Business Intelligence Tools. Computational Intelligence and Decision Making , 267-276.
[12]. "Welcome to Apache Hadoop!". hadoop.apache.org. Retrieved 2017-03-25.
[13]. "What is the Hadoop Distributed File System (HDFS)?". ibm.com. IBM. Retrieved 2017-03-25.
[14]. Malak, Michael (2014-09-19). "Data Locality: HPC vs. Hadoop vs. Spark". datascienceassn.org. Data Science Association. Retrieved 2017-4-4.
[15]. "Continuuity Raises $10 Million Series A Round to Ignite Big Data Application Development Within the Hadoop Ecosystem". finance.yahoo.com. Marketwired. 2012-11-14. Retrieved 2014-10-30.
[16]. Jaspersoft (2017), "Jaspersoft", http://www.jaspersoft.com/
[17]. Pentaho (2017), "Pentaho Open Source BI", Janeiro, http://www.pentaho.org/.Retrieved 2017-03-25.
[18]. SpagoBI (2017), "Spago Business Intelligence", http://www.spagoworld.org/.Retrieved 2017-03-25.
[19]. Vanilla (2017), "True Open Source BI Platform", http://vanilla-bi.com/.Retrieved 2017-03-25.
[20]. Jorge Bernardino (2011). Open Source Business Intelligence Platforms for Engineering Education. 1st world engineering education flash week, Lisbon 2011.
[21]. Bernardino, J., & Tereso, M. (2013). Business Intelligence Tools. Computational Intelligence and Decision Making , 267-276.
[22]. Victor M. P., Azeem M., Ali S., and Malka N. H., .Pentaho and Jaspersoft: A Comparative Study of Business Intelligence Open Source Tools Processing Big Data to Evaluate Performances. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016