Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

Empirical Investigation of Big Data Analytical Tools: Comparative Analysis

Harmeet Singh
Machine Learning Engineer
New Delhi, India
harmeet123@ gmail.com

Gurpreet Singh Matharu
Artificial Intelligence Research Scholar
EECS, Lassonde School of Engineering, York University
Toronto, Canada
mtech.gurpreet@gmail.com

Ashmeet Kaur Dardi
Artificial Intelligence Researcher
PhD (Computer Science)
Edmonton, Canada
adardi@gmail.com

Jaspreet Singh Matharu
Legal Analytics Enthusiast
Practicing Advocate, Delhi High Court
New Delhi, India
jaspreetwebs@gmail.com

Abstract—In this paper, we carry out an empirical investigation of the Big Data Analytical tools, Google BigQuery, Datameer Big Data Analytics, Alteryx Designer and Pentaho Big Data Analytics. Further, this paper provides for a comparative analysis among the most popular Big Data Analytical tools. The results of the undertaken industrial survey reveal higher adoption of Pentaho Big Data Analytics in present-day software industry as compared to other Big Data Analytical tools.

Keywords—Big Data, Analytical Tools, BigQuery, Datameer, Alteryx, Pentaho.

I. INTRODUCTION

The world is witnessing the generation of enormous amounts of data. The traditional data processing tools have become incompetent to process huge magnitude of data being produced over the industry. Consequently, all hopes are being placed upon the big data analytical tools. Big data is defined as huge amount of data that gathers over time and is difficult to analyze through conventional data processing tools. Big data ranging from scientific data to unstructured data posted on the Web to multi-structured data is experiencing massive generation in comparison to structured data. As per the Wikibon's Big Data Market Forecast [16], the big data market is expected to be valued at $50 billion till 2017.

Big data analytics is the need of the hour. It can significantly contribute in business generation and germinate an array of innovative business ideas. Big data analytics provides an opportunity to leverage the huge data available at the disposal of the organization to the advantage of their business, resulting in heightened business generation and better service delivery. Big data analytics refers to the process of examining huge amounts of different types of structured/unstructured data, in an effort to uncover hidden patterns, unknown correlations and other important information. Such information can provide competitive edge over business rivals and prove to be an elixir to the business, resulting in increased revenue generation, efficient decision making, strategic marketing and effective customer retention. It can enable companies make better business decisions by analyzing huge volumes of data that may remain untapped by traditional business data analysis.

This study provides a comparison among the most popular big data analytical tools, BigQuery, Datameer, Alteryx and Pentaho. The survey results would certainly assist the software practitioners in choosing among these big data analytical tools. The rest of this paper is structured as follows. Section II undertakes a literature review, followed by section III which discusses the most popular big data analytical tools. Section IV provides a comparative analysis among BigQuery, Datameer, Alteryx and Pentaho big data analytical tools, followed by Section V which presents the empirical survey results ending up with Section VI which provides the conclusions.

II. LITERATURE REVIEW

The findings of the survey [10] conducted by Vitria Technology, Inc. revealed that over 94% companies felt that being able to analyze big data was very crucial to their business. Over 41% respondents stated that taking action on analyzed data in mere seconds or minutes only could give them competitive advantage.

Another survey [5] conducted by Gigaspaces concludes that over 80% companies consider big data important to their business with 43% indicating it as mission critical. Further, this survey also found out that 43% of respondents are planning to

deploy some tool to cope up with big data analysis and while only 37% are presently using some form of big data analytical tool. This indicates towards a significant adoption of big data analytical tools across the industry.

In research work done by Udaigiri Chandrasekhar [2], he carries out a comparison between enterprise big data analytical tools and open source big data analytical tools based on several parameters including computing environment, data amounts that can be processed, ease of use, decision making capabilities, time consumption, and pricing norms. The author concludes that the selection of category of big data analytical tools depends upon the usage pattern and requirements of the company.

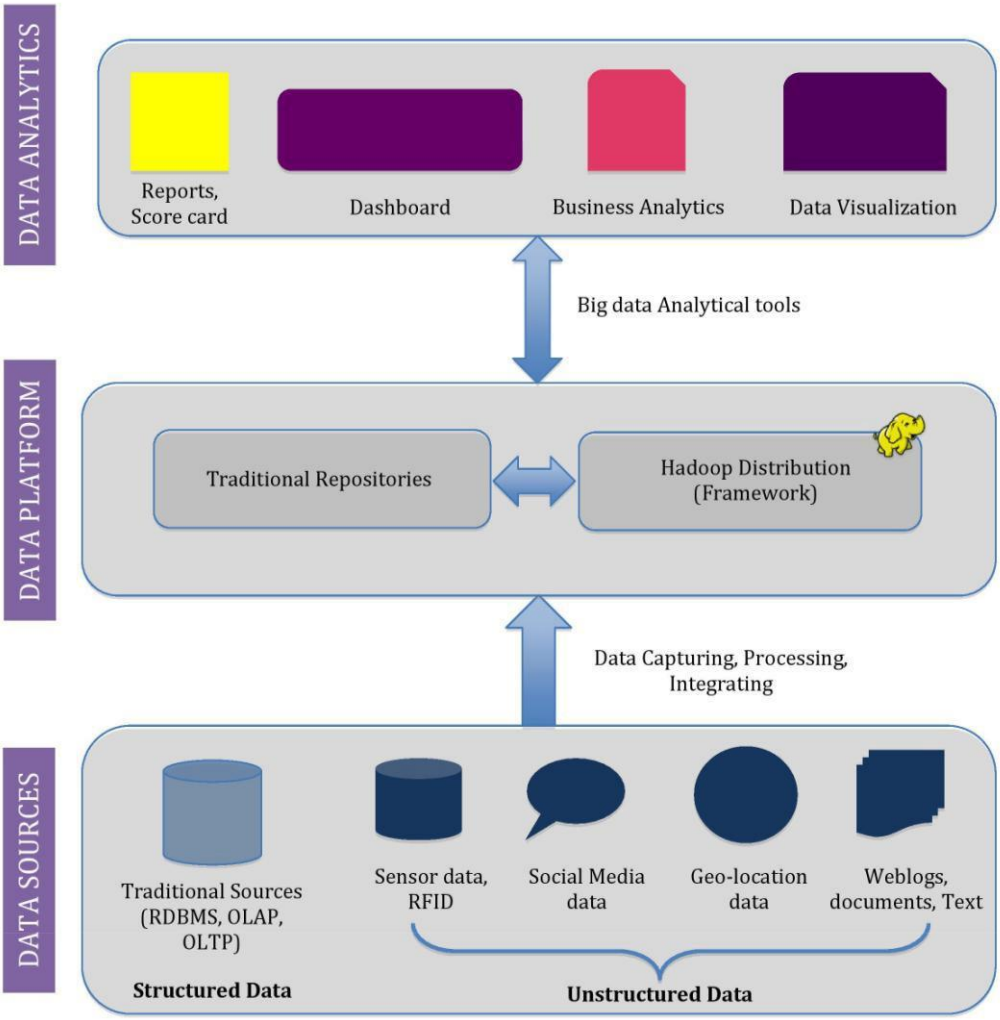

Fig. 1. Big Data Analytics Framework

III. BIG DATA ANALYTICAL TOOLS

A. *Google BigQuery*

Google BigQuery enables the users to run SQL queries over their unstructured data [1]. It allows executing SQL statements against huge datasets, running into billions of rows with breeze. It is a revolutionary tool capable of analyzing billions of data rows within seconds [2]. The distinguishing features of Google BigQuery are:

1) Analyses Large Datasets in Seconds: BigQuery involves executing super-fast, SQL-like queries against huge datasets, running into terabytes of data in time span of seconds by leveraging the processing power lying at the disposal of Google.

2) *Multiple Access Methods*: BigQuery can be accessed through BigQuery browser tool, Google Apps Script, Google Spreadsheets, or through BigQuery REST API.

3) Simplified Data loading: Data should be loaded into BigQuery before it can be used for analytics. Users may easily load their data through streaming, direct upload, or through Google Cloud Storage [1].

*4) Affordability***:** Google BigQuery processes the first 100 GB of data for free per month. They adhere to a simple and transparent pricing structure, where users need to pay only for the storage they need and the queries they run.

*5) Access Control***:** Users are free to share their datasets with individuals, groups, or even make it public.

6) Analytic functions: The power of BigQuery lies in its ability to interactively execute aggregate queries against terabytes of data. But, sometimes counts and averages are not sufficient, hence BigQuery provides various advanced functions including analytical functions for ranking results, exploring distributions and percentiles, and traversing results without the need for a self join.

7) Highly Scalable: High ingestion capacity of large amounts of data in mere minutes makes Google's Big highly scalable for any data volumes [19].

B. Datameer BigData Analytics

Datameer is a robust big data analytical application tool that offers data integration, analytics and visualization functionality, all built into one single application. The distinguishing features offered by Datameer BigData Analytics are:

1) Data Integration: Datameer allows gathering and integrating data from multiple resources into Hadoop using 20+ pre-built connectors [20].

2) REST API: This application comes handy with REST API that allows external applications to access Datameer data directly via code or command-line tools. Such programmatic access to raw data using APIs can surely streamline business decisions [4].

3) Spreadsheet-based Data Analysis: The data analytics functionality offered by Datameer has been built on a spreadsheet like interface that enables the users to analyze structured and unstructured data of any size by using built-in analytic functions and performs complex analytics, and all this is executed, without having to write any code [21].

4) Exporting Analytics: Datameer allows to export the analytics results to other data sources including NoSQL databases and RDBMS.

5) Instant Visualization: The 'flip side view' feature in Datameer lets the users flip any spreadsheet at any step in data analytics process, thus ensuring instant visualization of data in various different contexts.

6) Advanced Data Visualization: Once the users are done with data analysis, they can create visual representation of insights; create data visualization dashboards, time series, by using Datameer's wide variety of available widgets. Also, the visual insights can be viewed on all major platforms, including iPhone, iPad, MacOS and VMWare.

C. Alteryx Designer

Alteryx is another widely used data analytical tool across the industry. Alteryx is best known for its capability to rapidly process all sorts of sophisticated data. Alteryx offers an easy-to- use analytics platform for enterprises facilitating making critical decisions that drives their business strategy and organizational growth [22]. The main characteristics of Alteryx Designer are:

1) Data Blending: The data blending process involves integrating data from multiple sources in a single workflow that fuels definite business process and ensures improved decision making [9]. Alteryx seamlessly brings together distinct data from heterogeneous resources, such as consumer information from a traditional data source, with click stream or sensor data stored in Hadoop file system, which is necessary to gain insights when an organisation's infrastructure is insufficient to blend unrelated data together [6].

2) R Language: Data Analytics in Alteryx is deeply integrated with statistical language R (Revolution R Enterprise), due to which even the most sophisticated and complex data sets can be handled with ease [8].

3) Multi-threading: The multi- threading capability of Revolution R Enterprise embedded into Alteryx Analytics empowers data analysts to easily build analytical workflows and quickly perform advanced predictive analytics.

D. Pentaho BigData Analytics

Pentaho BigData Analytics tool offers a complete solution to blend, analyze and visualize the big data into insights within a single integrated platform. The unique features of Pentaho BigData Analytics are:

1) Blended Analytics: Pentaho big data analytical tool lets the users to blend all types of data. The data can be ingressed into Pentaho from a number of sources, including Hadoop, NoSQL databases, and relational databases. Powerful algorithms, built-in components and sophisticated data processing tools enable the users to uncover valuable insights that usually remain hidden on leveraging traditional data analysis [18].

2) Interactive Visualization: Pentaho big data analytical tool lets the users to draw interactive visualizations by leveraging the wide variety of built-in visualizations tools. The visualizations can also be viewed on Web-based responsive dashboards that are powered by rich interface, powerful graphics and a variety of filter controls. Pentaho's interactive visualizations, advanced analytical and predictive capabilities make it comprehensive, modern and complete big data analytical tool [17].

IV. COMPARISON AMONG BIG DATA ANALYTICAL TOOLS

Although several big data analytical tools are available in the market, but we have restricted our empirical investigation to some of the most popular big data analytical tools, resulting in the comparison as presented in Table I.

TABLE I. COMPARISON AMONG BIG DATA ANALYTICAL TOOLS

Big Data Analytical Tool	**Pentaho BigData Analytics**	**Alteryx Designer**	**Datameer BigData Analytics**	**Google BigQuery**
Vendor	Pentaho	Alteryx	Datameer	Google
Distinguished Feature	Provides blending of operational data sources with big data sources to create on demand analytical view [11]	Offers Analytics Gallery that allows single click consumption, sharing and publishing of analytic applications [12]	Offers Instant Visualization at every stage of analysis to gain insights [13]	Offers cloud storage with no charges for loading data and Google's own infrastructure
Offerings	Provides visualization, integration and analytics in single tool	Provides data Integration and Analytics	Integration, analytics and visualization	Cloud Server Storage, data analysis
Runs as	Standalone software	Standalone software	Standalone software	Cloud based service
Interface	Web Based interactive reporting view	Graphical user interface	Spreadsheet like interface [14]	Graphical Web Interface accessed through web Browser or cmd line tool
Data Import from Hadoop	Not all distributions are supported [15]	Supports all major distributions	Supports all Hadoop distributions	Not supported
Data Storage	On premise	On premise	Cloud as well as on premise	Google Cloud Storage
Pricing (Basic)	Customized license pricing	$3,995 Per User, Per Year	$2,999 for 10 Users.	Pay per use service, First 100 GB data processing free each month

V. SURVEY RESULTS

A mix of telephonic and email conversations were conducted with the representatives of 23 companies. As per the survey results, 10 companies were found to be leveraging Pentaho BigData Analytics, while 5 of them expressed their faith in Alteryx Designer, 4 of them are using Datameer BigData analytics and the remaining 4 companies have adopted Google BigQuery analytical tool. Table II enlists the companies and the big data analytical tools being leveraged by them and Fig.2 graphically depicts the relative number of companies using those big data analytical tools.

TABLE II. COMPANIES USING PENTAHO / ALTERYX / DATAMEER / GOOGLE BIGQUERY BIG DATA ANALYTICAL TOOLS

Big Data Analytical Tool	**Companies**
Pentaho BigData Analytics	• Lufthansa (www.lufthansa.com) • Kambi Sport (www.kambi.com)

	• 3CInteractive (www.3cinteractive.com) • BeachMint (www.beachmint.com) • Bywaters (www.bywaters.co.uk) • Delta Dental of Virginia (www.deltadentalva.com) • Edo Interactive (www.edointeractive.com) • Marketo (www.marketo.com) • Travian Games (www.traviangames.com) • Travel Tainment (www.traveltainment-group.com)
Alteryx Analytics	• The Boston Consulting Group (www.bcg.com) • Consumer Orbit (www.consumerorbit.com) • Valassis Communications (www.valassis.com) • Rentrak Corporation (www.rentrak.com) • Family Dollar (www.familydollar.com)
Datameer BigData Analytics	• Bank of America Corporation (www.bankofamerica.com) • American Airlines Inc. (www.aa.com) • American Express Company (www.americanexpress.com) • Comcast Corporation (www.comcast.com)
Google BigQuery	• Safari Books Online (www.safaribooksonline.com) • RedBus (www.redbus.in) • Boo-Box (www.boo-box.com) • Crystalloids (www.crystalloids.com)

VI. CONCLUSION

Big Data analytical tools provide an effective way to leverage the huge amounts of data being generated within the organizations to streamline their business strategies. The sudden increase in adoption of big data analytical tools across the industry proves their usefulness and effectiveness. Big data analytical tools result in effective decision making and innovative business planning. With our empirical study, we intend to contribute to the already existing body of knowledge about the comparative analysis of big data analytical tools.

Our survey results demonstrate higher adoption of Pentaho BigData Analytics in comparison to other big data analytical tools. The comparative analysis of four big data analytical tools, namely Datameer BigData Analytics, Alteryx Analytics, Datameer BigData Analytics and Google BigQuery conclude that although all the big data analytical tools provide the same functionalities of data analysis, integration and visualization, but they do differ on various parameters. Among all big data analytical tools, Pentaho BigData Analytics has the highest adoption whereas Alteryx Analytics is certainly picking up pace with data analysts starting to exploit it to their advantage, and Datameer BigData Analytics and Google BigQuery big data analytical tools are increasingly being explored by data analysts across the industry.

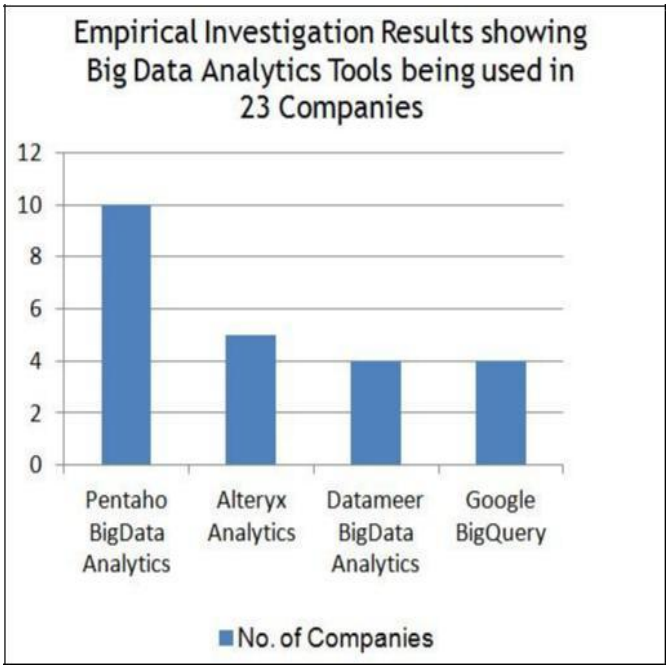

Fig. 2. Bar Graph depicting that Pentaho Big Data Analytics is being used more in comparison to other big data analytical tools

REFERENCES

[1] J. Tigani and S. Naidu in *Google BigQuery Analytics,* 1st ed. New Jersey, USA: John Wiley & Sons, 2014.

[2] U. Chandrasekhar, A. Reddy and R. Rath, "A Comparative Study of Enterprise and Open Source Big Data Analytical Tools," in *Proceedings of IEEE Conference on Information and Communication Technologie*s, JeJu Island, Korea, April 11-12, 2013, pp. 1-3. DOI=10.1109/CICT.2013.6558051

[3] Tech Republic, "Datameer 4.0 enables instant visualization in big data analytics," April 2014. Internet: http://www.techrepublic.com/article/datameer-4-0-enables-instant-visualization-in-big-data-analytics/

[4] Datameer, "Rapid big data integration". Internet: http://www.datameer.com/product/data-integration.html

[5] Gigaspaces, "Big Data Survey," 2012. Internet: http://www.gigaspaces.com/sites/default/files/product/BigDataSurvey_Report.pdf

[6] Alteryx, "Alteryx Designer". Internet: http://www.alteryx.com/products/alteryx-designer
[7] Teradata, "Alteryx". Internet: http://in.teradata.com/partners/Alteryx/?LangType=16393&LangSelect=true
[8] Revolution Analytics, "Alteryx and Revolution Analytics," *Data Sheet*. Internet: http://www.revolutionanalytics.com/sites/default/files/alteryx-revolution-analytics-ds.pdf
[9] Alteryx, "Alteryx: The Definitive Guide to Data Blending," http://pages.alteryx.com/rs/alteryx/images/ALT_WPDefGuideDataBlending-WithGraphics38.pdf
[10] Vitria, "Big Data, Streaming Analytics, And Breakthrough Operational Intelligence," Dec. 2013. Internet: http://www.vitria.com/pdf/Report-Big-Data-Analytics-Survey-2013.pdf
[11] Pentaho, "Big Data Blueprints". Internet:
[12] Alteryx, "Alteryx Analytics Gallery," *Data Sheet*. Internet: http://www.alteryx.com/sites/default/files/downloads/alt-gallery-ds-v3a.pdf
[13] Datameer, "Instant Visualizations in Every Step of Analysis". Internet: http://info.datameer.com/Web-Instant-Visualizations-in-Every-Step-of-Analysis-OnDemand.html
[14] Datameer, "End-to-end big data analytics". Internet: http://www.datameer.com/product/
[15] Pentaho, "Pentaho and Hadoop". Internet: http://www.pentaho.com/big-data-analytics/hadoop
[16] Wikibon, "Big Data Vendor Revenue and Market Forecast 2013-17," Feb. 2014. Internet: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
[17] Cloudera, "Cloudera and Pentaho". Internet: http://www.cloudera.com/content/dam/cloudera/Resources/PDF/solution-briefs/Cloudera-Pentaho-SolutionBrief.pdf
[18] Cio, "Pentaho Addresses Data Blending with Updated Business http://www.cio.com/article/2382507/big-data/pentaho-addresses-data-blending-with-updated-business-analytics-platform.html
[19] Saama Technologies, Inc., "How to Ingest Data into Google BigQuery using Talend for Big Data," *Technical Paper,* July 2013. Internet: http://www.saama.com/wp-content/uploads/2013/07/GoogleBlog_BigQueryBigData2.pdf
[20] Datameer, "Big Data Analytics," *Data Sheet,* 2012. Internet*:* http://www.datameer.com/pdf/Datameer_Big-Data-Analytics.pdf
[21] Datameer, "Retail: Customer Analysis," 2013. Internet: http://www.datameer.com/pdf/Datameer_Retail_Customer-Analytics.pdf
[22] Alteryx, "Teradata and Alteryx," *Data Sheet*, 2013. Internet: http://www.alteryx.com/sites/default/files/downloads/teradata-alteryx-joint-datasheet.pdf