A Survey: Software Tools Used In Big Data Analytics

[1]M. Sashi Kala* and [2]Dr. Nancy Jasmine Goldena
[1,2]*Assistant Professor, Department of Computer Science, Sarah Tucker College, TamilNadu, India

Abstract—Currently, organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to bemanaged and analyzed through traditional means. Every day, Google alone processes about 24 petabytes (or 24,000 terabytes) of data. Yetvery little of the information is formatted in the traditional rows and columns of conventional databases.Analyzing and working with Big Data could be very difficult using classical means like relational database management systems or desktop software packages for statistics and visualization. Instead, Big Data requires large clusters with hundreds or even thousands of computing nodes. This paper would highlight the software tools used successfully and widely for storage and processing of Big Data sets on clusters of commodity hardware. The primary purpose of this paper is to provide an in-depth analysis of different platforms available for performing big data analytics. The tools used for Extraction, Storage, Cleaning, Mining, Visualizing, Analyzing and Integrating are shed light on in detail.

Keywords: *Hadoop, HDInsight, Spark, Mozenda,* D3, Rapidminer, Orange, KNIME, Highchart, jHepwork

I. 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 is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.

Data analytics is primarily conducted in business-to-consumer (B2C) applications. Global organizations collect and analyze data associated with customers, business processes, market economics or practical experience. Data is categorized, stored and analyzed to study purchasing trends and patterns.

Evolving data facilitates thorough decision-making. For example, a social networking website collects data related to user preferences, community interests and segment according to specified criteria such as demographics, age or gender. Proper analysis reveals key user and customer trends and facilitates the social network's alignment of content, layout and overall strategy. Alternatively, more complex predictive and prescriptive modeling can help companies anticipate business opportunities and make decisions that affect profits in different areas.With predictive analytics, historical data sets are mined for patterns indicative of future situations and behaviors, while prescriptive analytics subsumes the results of predictive analytics to suggest actions that will best take advantage of the predicted scenarios.

Big data analytics tools are software products that support predictive and prescriptive analytics applications running on big data computing platforms -- typically, parallel processing systems based on clusters of commodity servers, scalable distributed storage and technologies such as Hadoop and NoSQL databases. The tools are designed to enable users to rapidly analyze large amounts of data, often within a real-time window. In addition, big data analytics tools provide the framework for using data mining techniques to analyze data, discover patterns, propose analytical models to recognize and react to identified patterns, and then enhance the performance of business processes by embedding the analytical models within the corresponding operational applications.

We focus on tools that meet the following criteria:

- They provide the analyst with advanced analytics algorithms and models.
- They're engineered to run on big data platforms such as Hadoop or specialty high-performance analytics systems.
- They're easily adaptable to use structured and unstructured data from multiple sources.
- Their performance is capable of scaling as more data is incorporated into analytical models.
- Their analytical models can be or already are integrated with data visualization and presentation tools.
- They can easily be integrated with other technologies.

In addition, the tools must incorporate essential characteristics and include integrated algorithms and methods supporting the typical suite of data mining techniques, including (but not limited to):

- **Clustering and segmentation:** divides a large collection of entities into smaller groups that exhibit some (potentially unanticipated) similarities. An example is analyzing a collection of customers to differentiate smaller segments for targeted marketing.
- **Classification:** is a process of organizing data into predefined classes based on attributes that are either pre-selected by an analyst or identified as a result of a clustering model. An example is using the segmentation model to determine into which segment a new customer would be categorized.
- **Regression:** which is used to discover relationships among a dependent variable and one or more independent variables, and helps determine how the dependent variable's values change in relation to the independent variable values. An example is using geographic location, mean income, average summer temperature and square footage to predict the future value of a property.
- **Association and item set mining:** looks for statistically relevant relationships among variables

in a large data set. For example, this could help direct call-center representatives to offer specific incentives based on the caller's customer segment, duration of relationship and type of complaint.

- **Similarity and correlation:** is used to inform undirected clustering algorithms. Similarity-scoring algorithms can be used to determine the similarity of entities placed in a candidate cluster.
- **Neural networks:** are used in undirected analysis for machine learning based on adaptive weighting and approximation.

This is just a subset of the types of analyses used for predictive and prescriptive analytics. In addition, different vendors are likely to provide a variety of algorithms supporting each of the different methods

II. TYPES OF DATA ANALYTICS

At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis.

Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

III. STEPS INVOLVED IN DATA ANALYTICS

The analytics process starts with the following:

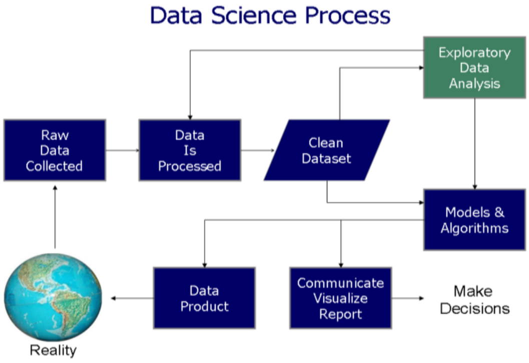
Data collection, in which data scientists identify the information they need for a particular analytics application and then work on their own or with data engineers and IT staffers to assemble it for use.

- Data from different source systems may need to be combined via data integration routines, transformed into a common format and loaded into an analytics system, such as a Hadoop cluster, NoSQL database or data warehouse. In other cases, the collection process may consist of pulling a relevant subset out of a stream of raw data that flows into, say, Hadoop and moving it to a separate partition in the system so it can be analyzed without affecting the overall data set.
- Once the data that's needed is in place, the next step is to find and fix data quality problems that could affect the accuracy of analytics applications. That includes running data profiling and data cleansing jobs to make sure that the information in a data set is consistent and that errors and duplicate entries are eliminated.
- Additional data preparation work is then done to manipulate and organize the data for the planned analytics use, and data governance policies are applied to ensure that the data hews to corporate standards and is being used properly.
- At that point, the data analytics work begins in earnest. A data scientist builds an analytical model, using predictive modeling tools or other analytics software and programming languages such as Python, Scala, R and SQL.
- The model is initially run against a partial data set to test its accuracy; typically, it's then revised and tested again, a process known as "training" the model that continues until it functions as intended. Finally, the model is run in production mode against the full data set, something that can be done once to address a specific information need or on an ongoing basis as the data is updated.

IV. TOOLS USED IN BIG DATA ANALYTICS

1. Store and Query /Analyse Data

The SQL Server Query Store feature provides you with insight on query plan choice and performance. The SQL Server Query Store feature provides you with insight on query plan choice and performance. It simplifies performance troubleshooting by helping you quickly find performance differences caused by query plan changes. Query Store automatically captures a history of queries, plans, and runtime statistics, and retains these for your review. It separates data by time windows so you can see database usage patterns and understand when query plan changes happened on the server.



a) Apache Hadoop

The Apache Hadoop is open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Apache Hadoop includes these modules:

- Hadoop Common: The most common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

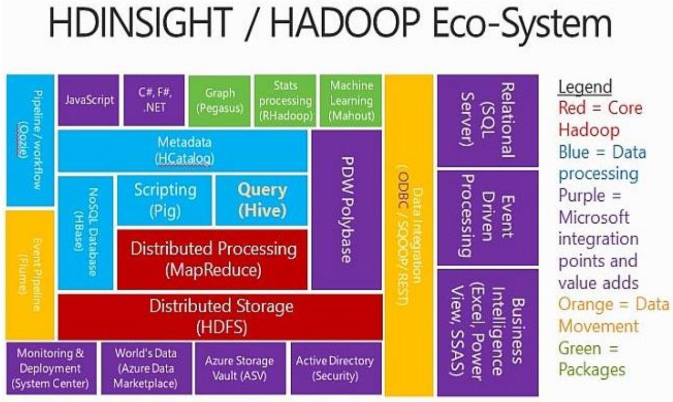
Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is not for the data beginner

b) Microsoft HDInsight

Azure HDInsight is the only fully-managed cloud Apache Hadoop offering that gives you optimized open-source analytic clusters for Spark, Hive, MapReduce, HBase, Storm, Kafka, and Microsoft R Server backed by a 99.9% SLA (service level agreement). Deploy these big data technologies and ISV (independent software vendor) applications as managed clusters with enterprise-level security and monitoring. Azure HDInsight is a cloud distribution of the Hadoop components from the Hortonworks Data Platform (HDP).

There are two flavors of HDInsight: Windows Azure HDInsight Service and Microsoft HDInsight Server for Windows (recently quietly killed but lives on in a different form)

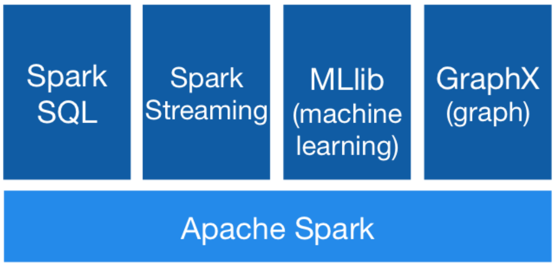
- Windows Azure HDInsight Service : is a service that deploys and provisions Apache Hadoop clusters in the Azure cloud, providing a software framework designed to manage, analyze and report on big data. It makes the HDFS/MapReduce software framework and related projects such as Pig, Sqoop and Hive available in a simpler, more scalable, and cost-efficient environment


HDINSIGHT / HADOOP Eco-System
Pipeline / workflow (Oozie)
JavaScript
C#, F#, .NET
Graph (Pegasus)
Stats processing (RHadoop)
Machine Learning (Mahout)
Metadata (HCatalog)
NoSQL Database (HBase)
Scripting (Pig)
Query (Hive)
Distributed Processing (MapReduce)
PDW Polybase
Event Pipeline (Flume)
Distributed Storage (HDFS)
Data Integration (ODBC / SQOOP/ REST)
Relational (SQL Server)
Event Driven Processing
Business Intelligence (Excel, Power View, SSAS)
Monitoring & Deployment (System Center)
World's Data (Azure Data Marketplace)
Azure Storage Vault (ASV)
Active Directory (Security)
Legend
Red = Core Hadoop
Blue = Data processing
Purple = Microsoft integration points and value adds
Orange = Data Movement
Green = Packages

- Microsoft HDInsight Server for Windows : was killed shortly after it was released but lives on in two flavors: Hortonworks Data Platform (HDP) and Microsoft's Parallel Data Warehouse (PDW). Both are on-premise solutions. With HDP, it includes core Hadoop (meaning the HDFS and MapReduce), plus Pig for MapReduce programming

c) Spark

Spark SQL is a new module in Apache Spark that integrates relational processing with Spark's functional programming API. Built on our experience with Shark, Spark SQL lets Spark programmers leverage the benefits of relational processing (e.g., declarative queries and optimized storage), and lets SQL users call complex analytics libraries in Spark (e.g., machine learning). Compared to previous systems, Spark SQL makes two main additions. First, it offers much tighter integration between relational and procedural processing, through a declarative DataFrame API that integrates with procedural Spark code. Second, it includes a highly extensible optimizer, Catalyst, built using features of the Scala programming language that makes it easy to add composable rules, control code generation, and define extension points. Using Catalyst, we have built a variety of features (e.g., schema inference for JSON, machine learning types, and query federation to external databases) tailored for the complex needs of modern data analysis. We see Spark SQL as an evolution of both SQL-on-Spark and of Spark itself, offering richer APIs and optimizations while keeping the benefits of the Spark programming model.Spark offers benefits such as automatic optimization, and letting users write complex pipelines that mix relational and complex analytics. It supports a wide range of features tailored to large-scale data analysis, including semi-structured data, query federation, and data types for machine learning. To enable these features, Spark SQL is based on an extensible optimizer called Catalyst that makes it easy to add optimization rules, data sources and data types by embedding into the Scala programming language. User feedback and benchmarks show that Spark SQL makes it significantly simpler and more efficient to write data pipelines that mix relational and procedural processing, while offering substantial speedups over previous SQL-on-Spark engines



2. Data Mining Software

a) Mozenda

This tool enables users to extract and manage Web data .Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, the data to be used in other applications or as intelligence.

There are two parts of Mozenda's scraper tool:

- Mozenda Web Console: It is a Web application that allows user to run agents, view & organize results, and export publish data extracted
- Agent Builder: It is a Windows application used to build data extraction project. Features:
  - Easy to use.
  - Platform independency.
  - Working place independence.

b) R

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians for developing statistical software and data analysis. According to Rexer's Annual Data Miner Survey in 2010, R has become the data mining tool used by more data miners (43%) than any other. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity. R is emerging as a defacto standard for computational statistics and predictive analytics. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- An effective data handling and storage facility.
- A suite of operators for calculations on arrays, in particular matrices.

- A large, coherent, integrated collection of intermediate tools for data analysis.
- Graphical facilities for data analysis and display either on-screen or on hardcopy.
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

c) *Orange*

Orange is all about data visualizations that help to uncover hidden data patterns provide intuition behind data analysis procedures or support communication between data scientists and domain experts. Visualization widgets include scatter plot, box plot and histogram, and model-specific visualizations like dendrogram, silhouette plot, and tree visualizations, just to mention a few. Much other visualization are available in add-ons and include visualizations of networks, word clouds, geographical maps, and more.

3. Data Visualization Software

a) D3

D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

D3 allows you to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. For example, you can use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive SVG bar chart with smooth transitions and interaction.

D3 is not a monolithic framework that seeks to provide every conceivable feature. Instead, D3 solves the crux of the problem: efficient manipulation of documents based on data. This avoids proprietary representation and affords extraordinary flexibility, exposing the full capabilities of web standards such as HTML, SVG, and CSS. With minimal overhead, D3 is extremely fast, supporting large datasets and dynamic behaviors for interaction and animation. D3's functional style allows code reuse through a diverse collection of official and community-developed modules.

b) Highcharts

Highcharts is a pure JavaScript based charting library meant to enhance web applications by adding interactive charting capability. Highcharts provides a wide variety of charts. For example, line charts, spline charts, area charts, bar charts, pie charts and so on. Highcharts is a pure JavaScript based charting library meant to enhance web applications by adding interactive charting capability. It supports a wide range of charts. Charts are drawn using SVG in standard browsers like Chrome, Firefox, Safari, Internet Explorer(IE). In legacy IE 6, VML is used to draw the graphics.

c) *jHepWork*

jHepWork is an environment for scientific computation, data analysis and data visualization for scientists, engineers and students. The program is fully multiplatform and integrated with the Jython (Python) scripting language. jHepWork can be used with several scripting languages for the Java platform, such as Jython, (the Python programming language), JRuby (the Ruby programming language) and BeanShell. This brings more power and simplicity for scientific computation. The programming can also be done in native Java. Symbolic calculations can be done using Matlab/Octave high-level interpreted language. The libraries include numerical and analytical calculations, linear algebra operations, equation solving algorithms.

4. Data Analytics Software

a) Rapidminer

Rapidminer is both a free open source and commercial product for text mining (content analysis).RapidMiner provides data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment. The data mining processes can be made up of arbitrarily nestable operators, described in XML files and created in RapidMiner's graphical user interface (GUI). RapidMiner is written in the Java programming language. It also integrates learning schemes and attribute evaluators of the Weka machine learning environment and statistical modelling schemes of the R-Project.

RapidMiner Studio's highlights are:

- A visual - code-free - environment, so no programming needed
- Available on all major operating systems and platforms
- Main function : Design of analysis processes
- Predictive analytics (with pre-made templates)
- Data loading
- Data transformation
- Data modeling
- Data visualization (with lots of visualizations)
- Extension API
- Lots of data sources : Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files, and more
- RapidMiner allows you to work with different types and sizes of data sources

b) KNIME

KNIME allows users to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later inspect the results, models, and interactive views. KNIME is written in Java and based on Eclipse and makes use of its extension mechanism to add plugins providing additional functionality. The core version already includes hundreds of modules for data integration (file I/O, database nodes supporting all common database management systems through JDBC), data transformation (filter, converter, combiner) as well as the commonly used methods for data analysis and visualization. With the free Report Designer extension, KNIME workflows can be used as data sets to create report templates that can be exported to document formats like doc, ppt, xls, pdf and others.

Other capabilities of KNIME are:

- KNIMEs core-architecture allows processing of large data volumes that are only limited by the available hard disk space (most other open source data analysis tools work in main memory and are therefore limited to the available RAM). E.g. KNIME allows analysis of 300 million customer addresses, 20 million cell images and 10 million molecular structures.

- Additional plugins allows the integration of methods for Text mining, Image mining, as well as time series analysis.
- KNIME integrates various other open-source projects, e.g. machine learning algorithms from Weka, the statistics package R project, as well as LIBSVM, JFreeChart, ImageJ, and the Chemistry Development Kit.

CONCLUSION

Big Data is going to continue growing during the next years and each data scientist will have to manage much more amount of data every year.This This data is going to more diverse, larger, faster and is becoming the new scientific data research and for business applications. Big data mining is new era which is help to discover knowledge. Big data analysis helps business people to make better decisions and researchers to identify new opportunities. This paper presents fundamental concepts of Big data like characteristics, sources, statistics, frameworks and technologies to handle big data

References

[1] Suman Arora, Dr.MadhuGoel, ―*Survey Paper on Scheduling in Hadoop*‖ International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[2] Puneet Singh Duggal ,Sanchita Paul ,― *Big Data Analysis:Challenges and Solutions*‖, *International Conference on Cloud, Big Data and Trust 2013, Nov 13-15.*

[3] Singh and Reddy, ‖A Survey on platforms for big data Analytics‖ Journal of Big Data 2014.

[4] Eckerson, W. (2011) ―Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations,‖ TDWI, September.

[5] Blog post: Thoran Rodrigues in Big Data Analytics, titled ―10 emerging technologies for Big Data‖, December 4, 2012

[6] J. Manyika, C. Michael, B. Brown et al., “Big data: The next frontier for innovation, competition, and productivity,” Tech. Rep., Mc Kinsey, May 2011.

[7] J. M. Wing, “Computational thinking and thinking about computing,” Philosophical Transactions of the Royal Society ofLondon A:Mathematical, Physical and Engineering Sciences, vol.366, no. 1881, pp. 3717–3725, 2008.

[8] J.Mervis, “Agencies rally to tackle big data,” Science, vol. 336, no.6077, p. 22, 2012.

[9] K. Douglas, “Infographic: big data brings marketing big numbers, ”2012, http:// www.marketingtechblog.com/ibm-big-datamarketing.

[10] S. SagirogluandD. Sinanc, “Big data: a review,” in Proceedings of the International Conference on Collaboration Technologies andSystems (CTS ’13), pp. 42–47, IEEE, San Diego, Calif, USA,May2013.

[11] Intel, “Big Data analaytics, ” 2012, p://www.intel.com/content/

[12] dam/www/public/us/en/documents/reports/data-insightspeer-research-report.pdf.

[13] https://azure.microsoft.com/en-us/solutions/big-data/

[14] https://hadoop.apache.org/Hadoop%3F

[15] http://spark.apache.org.

[16] https://en.wikipedia.org/wiki/Data_analysis

[17] http://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/

[18] https://docs.microsoft.com/en-us/sql/relational-databases/performance/monitoring-performance-by-using-the-query-store

[19] http://searchdatamanagement.techtarget.com/definition/data-analytics

[20] http://www.cs.waikato.ac.nz/ml/weka/

[21] https://orange.biolab.si