Journal of Management and Science ISSN: 2249-1260 | e-ISSN: 2250-1819

DOI:26524/jms.2017.24

Big data analytics steps and tools used in analytical process

Dr. V.Sasikala M.C.A., M.Phil., Ph.D., S.E.T
Assistant Professor , Department of Computer Science
DRBCCC Hindu College,pattabiram,Chennai-72

Introduction

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.[1]

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.

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis.

This paper reviews the steps followed in big data analytics and tools used in it. The steps involved in Big data analysis are Data Cleaning, Data Mining, Data Analysis, Data Visualization**,** Data Integration, Data Languages and Data Collection.

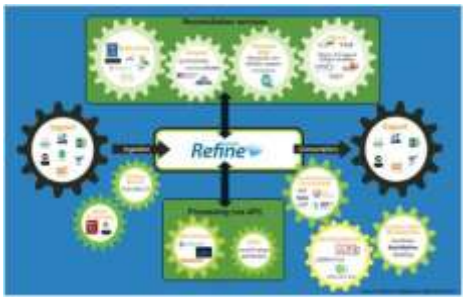
1 Step Followed In Big Data Analytics

1.1 Data Cleaning

Data scrubbing also called **data cleansing**, is the process of amending or removing **data** in a database that is incorrect, incomplete, improperly formatted, or duplicated. Data sets can come in all shapes and sizes, especially when we‘re getting it from the web. [2]

Tools Used in Data Cleaning

OpenRefine



OpenRefine is a data manipulation tool which cleans, reshapes and intelligently edit batch messy, and unstructured data. It is an open source tool and its code can be reused in other projects too. OpenRefine offers many features like faceting, clustering, editing cells, reconciling, extending web services, which helps to clean and transform data effectively. OpenRefine is easy as excel and powerful like access database. It makes many common tasks easy to do. It helps to analyze the data through filtering, faceting and converts the data into a more structured format.

Strengths of OpenRefine

1. OpenRefine is a desktop application. It opens in the browser as a Local Webserver. So, the data is safe and it doesn‘t get uploaded to the Google server.

2. It has facets which is used to filter the data into subsets and these clusters can be customized and organized into meaningful data.

3. It has a Browser based interface, and so can handle more data efficiently.

4. OpenRefine has a strong feature in extending data -- user can use it to find Meta Data and it can be used to correlate with it.

1.1.2 DataCleaner

The heart of Data Cleaner is a strong data profiling engine for discovering and analyzing the quality of our data. Find the patterns, missing values, character sets and other characteristics of our data values.[3]

Profiling is an essential activity of any Data Quality, Master Data Management or Data Governance program

Avoid operational issues and bad customer experiences by identifying if we have the same persons, companies and products registered multiple times in our databases.

- Profiles and analyzes our database within minutes!
- Access almost any datastore - Oracle, MySQL, PostgreSQL, MS SQL Server, MongoDB, CUBRID, CSV files, Excel spreadsheets, dbase and more
- Discover patterns in our textual data with the Pattern Finder
- Find out which values occur the most with the Value Distribution profile
- Cleanse our contact details with name and address validations
- Detect duplicates using fuzzy logic and configurable weights and thresholds
- Merge our duplicates and create a single version of the truth
- Write data back to relational databases, CSV files, Excel spreadsheets or MongoDB databases

1.2 Data Mining

Data mining is the process of discovering insights within a database as opposed to extracting data from web pages into databases. The aim of data mining is to make predictions and decisions on the data we have at hand.

Tools Used in Data Mining

1.2.1 RapidMiner

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the

machine learning process including data preparation, results visualization, validation and optimization. RapidMiner is developed on an open core model.[4]

RapidMiner functionality can be extended with additional plugins which are made available via RapidMiner Marketplace. The RapidMiner Marketplace provides a platform for developers to create data analysis algorithms and publish them to the community.[5]

With version 7.0, RapidMiner included updates to its getting started materials, an updated user interface, and improvements to its data preparation capabilities.

1.2.2 IBM SPSS Modeler

IBM SPSS Modeler is a data mining and text analytics software application from IBM. It is used to build predictive models and conduct other analytic tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming. One of its main aims from the outset was to get rid of unnecessary complexity in data transformations, and to make complex predictive models very easy to use. The first version incorporated decision trees (ID3), and neural networks, which could both be trained without underlying knowledge of how those techniques worked.

IBM SPSS Modeler was originally named Clementine by its creators, Integral Solutions Limited. This name continued for a while after SPSS's acquisition of the product. [6]

1.2.3 Oracle data mining

Oracle Data Miner is the graphical user interface for Oracle Data Mining. Oracle Data Miner provides wizards that guide us through the data preparation, data mining, model evaluation, and model scoring process. we can use the code generation feature of Oracle Data Miner to automatically generate PL/SQL code for the mining activities that we can perform.

Oracle Data Mining (ODM) embeds data mining within the Oracle database. ODM algorithms operate natively on relational tables or views, thus eliminating the need to extract and transfer data into standalone tools or specialized analytic servers. ODM's integrated architecture

results in a simpler, more reliable, and more efficient data management and analysis environment. Data mining tasks can run asynchronously and independently of any specific user interface as part of standard database processing pipelines and applications. Data analysts can mine the data in the database, build models and methodologies, and then turn those results and methodologies into full-fledged application components ready to be deployed in production environments. The benefits of the integration with the database cannot be emphasized enough when it comes to deploying models and scoring data in a production environment. ODM allows a user to take advantage of all aspects of Oracle's technology stack as part of an application. Also, fewer "moving parts" results in a simpler, more reliable, more powerful advanced business intelligence application.

ODM provides single-user multi-session access to models. ODM programs can run either asynchronously or synchronously in the Java interface. ODM programs using the PL/SQL interface run synchronously; to run PL/SQL asynchronously requires using the Oracle Scheduler. For a brief description of the ODM interfaces, see "Java and PL/SQL Interfaces".[7]

1.3 Data Analysis

While data mining is all about sifting through our data in search of previously unrecognized patterns, data analysis is about breaking that data down and assessing the impact of those patterns overtime. Analytics is about asking specific questions and finding the answers in data.

Tools Used in Data Analysis

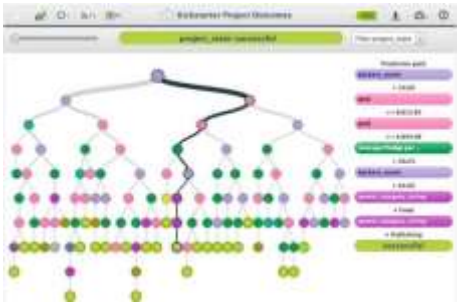
1.3.1 Qubole

Qubole delivers a more accessible way to perform big data analytics for data stored and growing in our AWS, Google and Microsoft cloud accounts. Qubole Data Service (QDS), serves as a unified interface for performing the myriad of use cases and workloads that a data driven organization will face ranging from ad hoc analysis, predictive analysis, machine learning, streaming and Map Reduce to name a few. Users without software development skills can

leverage the QDS workbench through our Smart Query interface without even knowing how to write a SQL query.

The QDS platform is built with extensibility as a core design principle. Developers may connect their applications to automatically drive queries using a number of QDS‘ language SDKs. Data analysts may prefer to compose queries directly for data preparation and ETL workloads, or interact with their BI/visualization tool of choice using one of the QDS ODBC connectors. QDS Data Engines are fully automated and optimized for the cloud. We are continually evaluating new and innovative open-source tools and engines as they gain customer traction. [8]

1.3.2 BigML



BigML, Inc. is a Machine learning company that provides software as a service (SAAS) for manipulating and analyzing data. The service can be used in production mode or development mode. Development mode is free but limited in the size of tasks that can be completed. Production mode is a paid mode and credits can be purchased ad hoc in blocks or on a subscription basis. This is a familiar pattern from other cloud based services like storage or compute servers. BigML provides three main modes to use the service:

- **Web Interface**: A slick web user interface that is fast and responsive. The web interface guides the analyst through the process of uploading data and making a descriptive or predictive model and evaluating the model or making predictions as needed.
- **Command Line Interface**: A command line tool called bigmler built upon the mature Python API for the service that allows more flexibility than the web interface such as the choice of making predictions against a constructed model locally or remotely, and performing tasks such as cross-validation to approximate model accuracy
- **API**: A RESTful API is provided that can be used directly via curl commands or via a wrapper in our favorite programming language.

BogML's products include : -

- **WhizzML** - It is a new domain-specific language for automating Machine Learning workflows, implementing high-level Machine Learning algorithms, and easily sharing them with others.

- **BigML.io** - It is a Machine Learning REST API to easily build, run, and bring predictive models to our project. we can use BigML.io for basic supervised and unsupervised machine learning tasks and also to create sophisticated machine learning pipelines.

- **BigMLer** - It is a command line tool, and it helps to automate the Machine Learning workflows in a single line.

- **The BigML PredictServer** - It is a dedicated machine image ideal to perform millions of predictions in real-time.

- **Flatline** - It is BigML‘s Lisp-like language that enables us to programmatically perform an array of data transformations, including filtering and new field generation. Flatliner is a handy code editor that helps to test the Flatline expressions.[9]

1.3.3 Statwing

Statwing is an easy-to-use statistical tool. Expert users work 5x faster in Statwing than they would in Excel or statistical tools like R or SPSS, and novice users can get as much insight out of their data as an expert data analyst. Statwing's modern, intuitive interface chooses statistical tests automatically, then reports results in plain English.

In every enterprise that uses analytics, there are a few power users who need the most advanced tools all of the time, and an army of casual users who need to do simple analysis now and then. For the latter group, cloud-based analytics make perfect sense; users get the tools they need when they need them, and the organization gets out of the business of licensing, hosting, distributing and maintaining infrequently used software.[10]

1.4 Data Visualization

Data visualization companies will make our data come to life. Part of the challenge for any data scientist is conveying the insights from that data to the rest of our company. For most of our colleagues, MySQL databases and spreadsheets aren‘t going to cut it. Visualizations are a bright and easy way to convey complex data insights.

Tools Used in Data Visualization

1.4.1 Tableau

There are many reasons why one should use tableau is, it is very easy to use and don‘t need to know programming of any sort, all we need is some data and tableau to create reports that are visually enchanting and which tells a story which we need to tell your managers or impress our professor in class. With its revolutionary drag and drop feature we can easily create stories or reports using just our mouse and a little imagination. All this is possible due to the revolutionary VizQL a visual query language

Advanced In-Memory Technology- The Data Engine Most of the data analytic software have a lot of fancy features but almost all of them fail when it comes to operating with large amounts of data, this is where the advanced in memory technology of tableau is a savior to all of those who need to get reports from ever increasing data. The tableau data engine is a breakthrough in-memory analytics database designed to overcome the limitations of existing databases and data silos. Capable of being run on ordinary computers, it leverages the complete memory hierarchy from disk to L1 cache. It shifts the curve between big data and fast analysis. And it puts that power into the hands of everyone. Ad-hoc analysis of massive data takes place in seconds. No fixed data model is required.[11]

1.4.2 Silk

Silk is a place to publish our data. Each Silk contains data on a specific topic. Anyone can explore a Silk and create interactive visualizations.

The features of Silk

Drag-and-Drop tools

Creating pages is fast and fun we can easily pull in images, videos, tables, maps, charts and more when building out our site.

Work together

Want to collaborate? Invite anyone to author or edit. It's perfect for teams who need to work together.

Keep info private

We have complete control over who can see your pages. Keep your Silk private or allow public

access.*Fast data importer*

Use it to convert our spreadsheets into Silk pages. our information will be tagged and viewable in minutes.

Auto Magical map creation

Silk automatically convert addresses into an online map that our users can navigate with ease. Heat maps, too!

Silk is for anyone and everyone

Artists, musicians, non-profits, museums, teachers, companies, sports fans, entertainment bloggers, statistical experts, and venture capitalists all use Silk.[12]

1.4.3 Chartio

Chartio saves our team over 30 hours each week, helping to automate routine analyses and enabling us to spend time gathering deeper insights.Easily share dashboards via scheduled emails or reports without having to install software. Quickly view historical data and track corporate metrics with our Snapshots feature**.** Customize and brand our dashboard for improved and consistent visualizations.

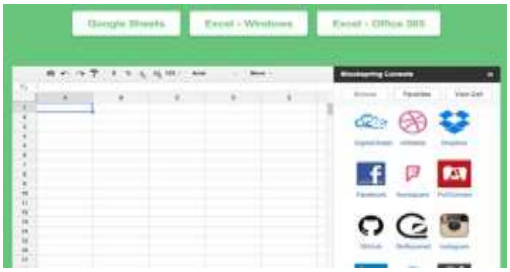
Out-of-the-box connections to our data sources from Amazon Redshift to CSV files, so we can start exploring data immediately. We speak our language. Because we're built on SQL there‘s no need to learn a proprietary language. Deploy our cloud-based application with a few clicks; no hardware required.[13]

1.5 Data Integration

Data integration platforms are the glue between each program. If we want to connect the data we‘ve extracted using Import.io with Twitter or we want to share on Facebook the visualisation we‘ve made with Tableau or Silk automatically, then the integration services below are the tools for us.

Tools used for Data Integration

1.5.1 Blockspring



Blockspring makes it easy to bring data from external APIs into one place. It is a unique program in the way that they harness all of the power of services such as Zapier in familiar platforms such as Excel and Google Sheets. We can connect to a whole host of 3rd party programs by simply writing a Google Sheet formula.

1.5.2 Pentaho

Pentaho is a Business Intelligence (BI) software company that offers Pentaho Business Analytics, a suite of open source products which provide data integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities.

The Pentaho suite consists of two offerings, an enterprise and community edition. The enterprise edition contains extra features not found in the community edition. The enterprise edition is obtained through an annual subscription and includes extra support services. Pentaho's core offering is frequently enhanced by add-on products, usually in the form of plug-ins, from the company itself and also the broader community of users and enthusiasts. The table below summarizes the most popular products and plug-ins in the Pentaho ecosystem.[14]

1.6 Data Languages

There will be times in our data career when a tool simply won‘t cut it. While today‘s tools are becoming more powerful and easier to use, sometimes it is just better to code it ourself. Even if we‘re not a programmer, understanding the basics of how these languages work will give us a better understanding of how many of these tools function and how best to use them.

1.6.1 R –Programming Language

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing.[5] The R language is widely used among statisticians and data miners for developingstatistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.[8]

R is a GNU package. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.[15]

1.6.2 Python

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy which emphasizes code readability and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java.[22][23] The language provides constructs intended to enable writing clear programs on both a small and large scale.

Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library.

Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems. CPython, thereference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation.[16]

1.7 Data Collection

Before we can store, analyze or visualize our data, we‘ve got to have some. Data extraction is all about taking something that is unstructured, like a webpage, and turning it into a structured table. Once we‘ve got it structured, we can manipulate it in all sorts of ways, using the tools we‘ve covered, to find insights.

1.7.1 Import.io

import.io is a web-based platform for extracting data from websites without writing any code. The tool allows people to converted unstructured web data into a structured format for use in Machine Learning, Artificial Intelligence, Retail Price Monitoring, Store Locators as well as academic and other research. It is also used extensively by investigative journalists.

Users enter a URL and the app attempts to automatically extract the data that it thinks we need, if the automatic extraction does not provide exactly what we need, a point and click interface allows us to "train" the app what to extract. The data that users collect is stored on Import.io‘s cloud servers and can be downloaded as CSV, Excel, Google Sheets, JSON or accessed via API. Users can easily integrate live web data into their own applications or third party analytics and visualization software. Thousands of data sources can be extracted simultaneously.

- Auto-extraction - Automatically extract data from web pages into a structured dataset
- Extractor builder - Point and click to build extractors
- Authentication - Extract data from behind a login/password
- Scheduler - Schedule extractors to run exactly when we need them to
- Online datastore - Use the SaaS platform to store data that is extracted
- Throughput - Fast, parallelized data acquisition distributed automatically by scalable cloud architecture
- Uptime - High availability for high volume usage
- Integrations - Integrations with Google Sheets, Excel, Tableau and many others. Generate example code to integrate with our own data sources in the language of your choice[17]

References:

1) Big data - Wikipedia
2) OpenRefine.org
3) wikipedia.org/wiki/Data cleansing
4) Wikipedia// rapidminer
5) https://rapidminer.com
6) wikipedia.org/wiki/SPSS_Modeler
7) Oracle Data Mining
8) Qubole wiki
9) cloudacademy.com/blog/bigml-machine-learning
10) www.qualtrics.com/statwing
11) Tableau Software - Wikipedia
12) https://www.statsilk.com/
13) data-visualization-with-chartio
14) Pentaho - Wikipedia
15) https://www.programiz.com/r-programming
16) www.python.org
17) Import.io - Wikipedia

Conclusion

In this paper I have discussed about the steps involved in Big data analysis, the steps to be followed for the analysis are Data Cleaning, Data Mining, Data Analysis, Data Visualization**,** Data Integration, Data Languages and Data Collection. For the above steps in data analysis the software‘s used for are also discussed.

*****