BOOK REVIEW

Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data

EMC Educational Services,
Key contributors: Dietrich, D., Helle, B., Yang, B
John Wiley & Sons, Inc., 2015
ISBN: 978-1-118-87613-8, 410 pp.

Innovative insights, improved understanding of problems, and opportunities to predict and shape the future are the main facets of the enormous potential in Big Data. In this context, Big Data analytics integrate structured and unstructured data with real time feeds and queries to enable a new path to innovation, to investigation with potentially deeper insights. *EMC Educational Services* proposes this book to individuals and organizations with a view to helping them benefit from this potential. The book provides *a practitioner's approach* to some of the key techniques and tools used in Big Data analytics that could assist multiple stakeholders – business and data analysts, database professionals and managers of business intelligence, Big Data groups looking to enrich their analytical skills – to become active contributors to Big Data analytics projects. It is also useful for university graduates interested in data science as a career field. This book prepares for the certification of Data Science Associate, EMC Certified Professional (EMCDSA).

The book is structured in twelve chapters. Each chapter starts with a list of the addressed key concepts, its objectives and their logical connection with previous chapters' contribution. At the end of each chapter a summary of its main conclusions and references to subsequent chapters that are based on its content are provided, as well as a list of exercises and the bibliography.

Chapter 1 – "*Introduction to Big Data Analytics*" explains several key concepts of the domain to clarify the meaning of Big Data, why advanced analytics are needed, how *Data Science* differs from *Business Intelligence*, what new roles are needed for the Big Data ecosystem with the focus on data scientist. Section 1.1 – "*Big Data Overview*" is focused on data characteristics and structures. Huge volume, complexity of data types and structures, speed of new data creation and growth which corresponds to well-known *3 Vs (volume, variety, and velocity*) are explained. Social media and genetic sequencing are presented as relevant examples of fastest-growing sources of Big Data. 80-90% of the data growth is coming from non-structured data types: *semi-structured data* (textual data with a discernable pattern that enables parsing, like XML data files), *quasi-structured data* (textual data with erratic data formats like web clickstream data), and *unstructured data* (data with no inherent structure, like text documents, PDFs, images, and video). From the analyst perspective these four types of data (structured and non-structured) are covered by three types of data repositories: *spreadsheets and datamarts* (low volume data bases), *data warehouses* ("data base administrator-owned" centralized data containers in a purpose-built space), and *analytic sandbox* ("analyst-owned" data assets gathered from multiple sources and technologies). Section 1.2 – "*State of the Practice in Analytics*" addresses the following topics: business drivers for advanced analytics (optimize business operations, identify business risks, predict new business opportunities, comply with legal and regulatory requirements), Business Intelligence versus Data Science (Predictive Analytics and Data Mining), current Analytical architecture (data sources, departmental and enterprise data warehouses, applications for Business Intelligence and reporting, downstream analytics for Data Science users), drivers of Big Data (data deluge generated by new types of applications and specific analysis requirements), emerging Big Data ecosystems

and a new approach to Analytics (built on four web interconnected main groups of players: data devices, data collectors, data aggregators, and data users and buyers). In Section 1.3 – "*Key roles for the new Big Data ecosystem*" the key roles of the new Big Data ecosystem are detailed (technology and data enablers, Data savvy professionals, Data scientists). Three recurring sets of activities and five main sets of skills and behavioral characteristics specific to Data scientists are described next. This chapter ends with providing three examples of Big Data Analytics solutions in retail, IT infrastructure and social media.

Chapter 2 – "*Data Analytics Lifecycle*" is based on the idea that Data Science projects differ from most traditional BI projects in that they are more exploratory in nature. Therefore it is critical to have a process to govern them and to ensure that the participants are thorough and rigorous in their approach, yet not so rigid that the process impedes exploration. The chapter presents *an analytic project life cycle* designed for the particular characteristics and challenges of hypothesis-driven analysis with Big Data. In Section 2.1 – "*Data Analytics Life Overview*" key roles for a successful analytics project are identified (business user, project sponsor, project manager, business intelligence analyst, database administrator, data engineer, data scientist) and an overview of the main phases of the data analytics lifecycle is provided. The remaining part of this chapter details these phases as follows: phase 1 – *Discovery* (learn the business domain, assess available project resources, frame the business problem as an analytics challenge, formulate initial hypotheses to test), phase 2 – *Data preparation* (prepare the data to work with it and analyze by extract, load, and transform (ELT) or extract, transform and load (ETL) operations), phase 3 – *Model planning* (determine the methods, techniques and workflow to follow for subsequent model building phase, explore the data, select the most suitable models), phase 4 – *Model building* (develop datasets for testing, execute models, evaluate existing tools for running the models), phase 5 – *Communicate results* (evaluate the project results together with major stakeholders, identify key findings, prepare their presentation), and phase 6 – *Operationalize* (deliver final reports and technical documents, optionally run a pilot project to implement the models in a production environment). The cycle is exemplified by the case study on GINA (Global Innovation Network and Analysis) to analyze innovation data at EMC.

Chapter 3 – "*Review of Basic Data Analytics Methods using R*" examines fundamental statistical techniques in the context of the open-source R analytic software environment, highlights the importance of exploratory data analysis via visualization and reviews the key notions of hypothesis development and testing. *The language R* was selected to illustrate analytical tasks and models due to its popularity and versatility. Section 3.1 – "*Introduction to R*" provides an overview of the basic functionality of R: graphical user interfaces, data import and export, attribute and data types, generate descriptive statistics. Section 3.2 – "*Exploratory Data Analysis*" underlines the advantages of visualization as succinct holistic view of data that may be difficult to grasp from the numbers and summaries alone: visualization of datasets before analysis, detect dirty data in the exploration phase, display data to help explain the underlying distributions of a single variable or the relationship between two or more variables, visualization for data exploration versus visualization for data presentation. Section 3.3. – "*Statistical methods for evaluation*" emphasizes the crucial role of statistics during the initial data exploration and data preparation, model building, evaluation of the final models, and assessment of how the new models improve the situation when deployed in the field. The following statistical tools are presented: hypothesis testing, difference of means, Wilcoxson rank-sum test, two types of errors (in connection with hypothesis test), effect size, analysis of variance (ANOVA).

Chapters 4 to 9 are built upon the introduction to R presented in Chapter 3 and discuss a range of *advanced analytical theory and methods*: clustering, association rules, regression, classification, time series analysis, and text analysis, respectively.

Clustering analysis groups similar objects based on the objects' attributes.

Chapter 4 is focused on the *k-means analytical technique* that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the centre of k groups.

The presentation is structured on use cases (image processing, medical investigations, customer segmentation), overview of the methods, determining the number of clusters, diagnostics, reasons to choose and cautions. If k-means does not appear to be an appropriate clustering technique for a given dataset, then *alternative techniques* such as k-moedes or PAM (partitioning around medoids) should be considered. Clustering can be used with other analytical techniques such as regression.

Association rules represent an unsupervised learning method used to discover interesting relationships hidden in a large dataset. Association rules are commonly used for mining transactions in databases.

Chapter 5 starts with an overview of this method and than describes *Apriori*, one of the earliest and most fundamental algorithms for generating association rules. To illustrate the application of association rules the market basket analysis concept is introduced. The grocery store example is used to walk through the steps of and generate frequent k-item sets and useful rules for downstream analysis and visualization. Finally, the chapter formulates some pros and cons of the *Apriori* algorithm and highlights a few methods to improve its efficiency.

In general, *regression analysis* attempts to explain the influence that a set of input or independent variables has on the outcome of another variable of interest, called dependent variable.

Chapter 6 examines two regression techniques, linear regression and logistic regression, and explains when one technique is more appropriate than the other. For *the linear regression* the key assumption is that the relationship between an input variable and the outcome variable is linear. The model description includes linear regression model with normally distributed errors, use of categorical variables, confidence intervals on the parameters, confident interval on the expected outcome, prediction interval on the particular outcome, tools and techniques to validate a fitted linear regression model. *The logistic regression* is used when the outcome variable is categorical in nature (true/false, pass/fail, yes/no): patient's successful response to a specific medical treatment or procedure, the probability that credit applicant will default on loan, the probability of experiencing a malfunction of failure in engineering. The chapter outlines the considerable care that must be taken in performing and interpreting a regression analysis and the activities that a data scientist has to do in this respect: determine the best input variables and their relationship to the outcome variable, understand the model assumptions and their impact on the modelling results, transform the variables, as appropriate, to achieve adherence to the model assumptions, decide on the best choice between building one comprehensive model or many models on partitions of the data.

Classification is widely used for prediction purposes. Most classification methods are supervised, in that they start with a training set of pre-labelled observations to learn how likely the attributes of these observations may contribute to the classification of future unlabelled observations.

Chapter 7 focuses on two fundamental classification methods: decision trees and naïve Bayes. A *decision tree* (also called prediction tree) uses a tree structure to specify sequences of decisions and consequences. The input variables of a decision tree can be categorical of continuous. There are two varieties of decision trees: classification trees, usually applied in case of categorical (often binary) output variables, and regression trees, which can apply to output variables that are numeric or continuous. This section of Chapter 7 includes: an overview of a decision tree, the general algorithm, decision tree algorithms, evaluating a decision tree, decision tree in R. *The Naïve Bayes* is a probabilistic classification method based on the Bayes theorem which gives the relationship between the probabilities of two events and their conditional probabilities. The input variables are generally categorical, but continuous variables can also be accepted. The description includes the Bayes' theorem, Naïve Bayes classifier, smoothing, diagnostics, Naïve Bayes in R. These classifiers along with logistic regression (Chapter 6) are often used for the classification of data. Recommendations for choosing a suitable classifier depending on characteristics of input data are also provided. Finally, some additional classification methods are mentioned: bagging, boosting, and random forest that use multiple models to obtain better predictive performance; support vector

machines that combine linear models with instant-based learning techniques.

Time series analysis attempts to model the underlying structure of observations taken over time. It implicitly addresses the case in which any particular observation is somewhat dependent on prior observations. The goals are to identify and model the structure of the time series and to forecast future values in the time series. The application areas include finance, economics, biology, engineering, retail and manufacturing.

Chapter 8 examines this topic and its applications. The presentation starts with *time series components* (trend, seasonality, cyclic, random) and Box-Jenkins methodology. The main focus of the chapter is on *ARIMA (Auto-regressive Integrated Moving Average)* model: autocorrelation function, auto-regressive models, moving average models, building and evaluating an ARIMA model, reasons to choose and cautions. *Additional methods* are also mentioned: spectral analysis, Kalman filtering, ARMAX, GARCH, multi-variable time series analysis.

Text analysis refers to the representation, processing, and modelling of the textual data to derive useful insights. An important component of the text analysis is the text mining, the process of discovering relationships and interesting patterns in large text collections. **Chapter 9** presents *the main text analysis steps*: parsing, search and retrieval, text mining. Then, the *text analysis process* is detailed: collecting raw text, representing text, using TFIDF (Text Frequency-Inverse Document Frequency), categorizing document by topics, determining sentiments, gaining greater insights. Considering the Data Analytic Lifecycle, the most time consuming part of this process is not performing the statistics or implementing algorithms, but formulating the problem and getting and preparing the data.

Chapter 10 – "*Advanced Analytics – Technology and Tools: MapReduce and Hadoop*" presents some key technologies and tools related to the Apache Hadoop software library. The chapter starts with *presentation of analytics for unstructured data*: the MapReduce paradigm, offering the means to break a large task into smaller tasks, to run them in parallel and consolidate the output, and the Apache Hadoop which includes an implementation of MapReduce. Hadoop distributed file system (HDFS), structuring the MapReduce job in Hadoop, additional considerations in structuring a MapReduce job, developing and executing a Hadoop MapReduce program, Yet Another Resource Negotiator (YARN) are detailed. In the next section, the *Hadoop ecosystem* is presented including the main Hadoop-related Apache projects: Pig (high-level data-flow programming language), Hive (providing SQL like access), HBase (real time reads and writes) and Mahout (analytical tools). The final section discusses *four major categories of NoSQL* (Not only Structured Query Language): key/value stores (e.g. Redis), document stores when the value of a key/value pair is a file and the file itself is self-describing (CouchDB, MongoDB), column family stores for sparse datasets (Cassandra, Hbase), graph databases to store items and relationship between them (FlockDB, Neo4j).

Chapter 11 – "*Advanced Analytics – Technology and Tools: In-Database Analytics*" deals with the processing of data within its repository. Advantages of in-database analytics include no need for movement of the data into an analytic tool and almost real time results. The chapter is structured on: *SQL essentials* (Joins, set operations, grouping extensions), *In-Database Text Analysis*, and *Advanced SQL* (Window function, user-defined functions and aggregates, ordered aggregates, MADlib). A typical SQL query is commonly associated with structured data, but SQL tables often contain unstructured data (text content). Regular expressions and related functions can be used in SQL to examine and restructure such data for further analysis. Window functions are used in complex SQL queries to supply computed values such as ranks and rolling averages. It is possible to process the data within a database and extract the results into an analytical tool such as R. Also external libraries such as MADlib can be utilized by SQL to conduct statistical analyses within a database.

Chapter 12 – "*The Endgame, or Putting It All Together*" focuses on the final phase of the Data Analytics Lifecycle: *communicating and operationalizing an analytics project* (key outputs of each of the main stakeholders), *creating the final deliverables* (developing core

materials for multiple audiences, main findings, the approach, model description, recommendations, additional tips on final presentation, providing technical specification and code), *data visualization basics* (key points supported with data, various types of graphs, common representation methods, how to clean up a graphic, additional considerations).

Overall the book may be considered a useful primer: it covers a wide range of topics, it is clear and informative.

Reviewed by:

Gabriel NEAGU, PhD

Senior Researcher

National Institute for R&D in Informatics – ICI-Bucharest

8-10 Averescu Blvd.

011455 Bucharest 1

ROMANIA