Information Processing and Management 000 (2017) 1–11

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Human resources for Big Data professions: A systematic classification of job roles and required skill sets

Andrea De Mauro[a,*], Marco Greco[b], Michele Grimaldi[b], Paavo Ritala[c]

[a] Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico, 1, 00133, Rome, Italy
[b] Department of Civil and Mechanical Engineering, University of Cassino and the Southern Lazio, Via G. Di Biasio, 43, 03043, Cassino FR, Italy
[c] School of Business and Management, Lappeenranta University of Technology, Finland

ARTICLE INFO

*Article history:*
Received 1 January 2017
Revised 7 May 2017
Accepted 23 May 2017
Available online xxx

*Keywords:*
Big Data
Business intelligence
Human resources management
Machine learning
Topic modeling

ABSTRACT

The rapid expansion of Big Data Analytics is forcing companies to rethink their Human Resource (HR) needs. However, at the same time, it is unclear which types of job roles and skills constitute this area. To this end, this study pursues to drive clarity across the heterogeneous nature of skills required in Big Data professions, by analyzing a large amount of real-world job posts published online. More precisely we: 1) identify four Big Data 'job families'; 2) recognize nine homogeneous groups of Big Data skills (skill sets) that are being demanded by companies; 3) characterize each job family with the appropriate level of competence required within each Big Data skill set. We propose a novel, semi-automated, fully replicable, analytical methodology based on a combination of machine learning algorithms and expert judgement. Our analysis leverages a significant amount of online job posts, obtained through web scraping, to generate an intelligible classification of job roles and skill sets. The results can support business leaders and HR managers in establishing clear strategies for the acquisition and the development of the right skills needed to leverage Big Data at best. Moreover, the structured classification of job families and skill sets will help establish a common dictionary to be used by HR recruiters and education providers, so that supply and demand can more effectively meet in the job marketplace.


## 1. Introduction

The phenomenon of Big Data – defined as those "*information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value*" (De Mauro, Greco, & Grimaldi, 2016) - has reached a prominent position on the agendas of business managers around the world. Multiple previous studies have proven the crucial role of human capital in the success of companies, especially those characterized by a high degree of technology intensity (Colombo & Grilli, 2010; Delgado-Verde, Martín-De Castro, & Amores-Salvadó, 2016; Morales-Alonso, Pablo-Lerchundi, & Núez-Del-Río, 2016; Saenz, Aramburu, Buenchea, Vanhala, & Ritala, 2017; Siepel, Cowling, & Coad, 2017). Consequently, firms need to quickly secure the appropriate competencies in the area of Big Data: such a race for acquiring the right talent does not seem to slow down while the labor market is unable to cope with an exponentially increasing

* Corresponding author.
*E-mail addresses:* andrea.de.mauro@uniroma2.it (A. De Mauro), m.greco@unicas.it (M. Greco), m.grimaldi@unicas.it (M. Grimaldi), paavo.ritala@lut.fi (P. Ritala).

http://dx.doi.org/10.1016/j.ipm.2017.05.004
0306-4573/

demand. Emerging Big Data job roles are making a rapid entry into the list of openings in the hands of corporate recruiters. However, the description of skills and responsibilities of Data Analytics jobs is often nebulous and companies tend to rely on subjective interpretations of their organizational needs. For instance, we have recently witnessed the establishment of the nearly-mythological role of Data Scientist, a professional able to individually cope with a company's most analytical necessities. This simplistic vision clearly neglects the complexity of all those varied and specific skills that are required to gather information, organize it and transform it into insights that can produce an economical advantage. In fact, there is a clear research gap regarding the formal definition of the most prominent Big Data jobs and of the required educational needs (Miller, 2014; Song & Zhu, 2015).

In this study, building on our previous work (De Mauro, Greco, Grimaldi, & Nobili, 2016), we aim to bridge the above-mentioned gap in the current knowledge by bringing clarity over this vital aspect of Big Data, i.e. its professional workforce and the most required skill sets. In particular, we provide a data-based description of job roles and skills that companies need in order to make use of Big Data. This enriches the literature by offering a structured framework for further research on competency requirements in this business-relevant field. The results also contribute to practice by providing an overarching understanding of what constitutes the contemporary professions around Big Data in organizations, helping HR and other managers to search for better recruitments and develop human capital towards desired directions.

The main data input of this study was a mass-download of a considerable number of online job posts, obtained by means of web-scraping techniques. Web-scraping consists in systematically collecting web pages through computer software. In our case, we have acquired more than 2.700 job posts which contained the keywords 'Big Data' in either the title or the description. We have then applied a set of text mining and classification algorithms in order to recognize which skills are required within each job type and to which extent. The final results confirm that 'Data Scientist' is an umbrella term that loosely describes the complex set of interconnected skills required by companies exploiting Big Data Analytics. Business Managers and HR professionals can use our 'job family vs. skill set' classification when designing job posts and when assessing the overall sufficiency of a company's human capital with regard to Big Data.

The paper is organized as follows: in Section 2, we present previous works on Big Data and Data Scientists, stripping away the many myths related to Data Science as a discipline. In Section 3 we describe the 4-step methodology we have used to acquire the job posts and systematically analyze their content, while in Section 4 we discuss the obtained results and provide a description of the job families we identified and their related skill sets. Section 5 summarizes our conclusions and suggests future extensions to the current work.

## 2. Related work

### 2.1. Big Data

The term 'Big Data Analytics' has become popular within IT communities and scholar research as of 2011 (Gandomi & Haider, 2014). Its meaning is the result of a disorganized evolution and the merge of several more traditional concepts such as: 'Very Large Databases' and 'Data Mining'. The vagueness of the concept of Big Data has resulted in a proliferation of multiple, sometimes contradictory definitions (Ward & Barker, 2013; Ylijoki & Porras, 2016). However, its essential characteristics (Information, Technology, Methods and Impact) have been identified and provide a conceptual framework for comprehending its overall meaning (De Mauro, Greco, & Grimaldi, 2016; van Altena, Moerland, Zwinderman, & Olabarriaga, 2016):

- **Information**: Our society is witnessing an unprecedented growth in information availability. In particular, as noticed by Hilbert (Hilbert, 2016), over the last two decades we have seen an exponential increase of information flow, stock and computational power. The characteristics of information are also changing quickly: data is now more 'personal', meaning that its majority is deemed to be created and consumed directly by human beings, as they interact with other individuals and machines through their personal devices. Data is now also more varied in type than it was in the past: in fact, traditional numeric datasets are becoming a small portion of the entire digital universe, which is acquiring more unstructured data types, such as audio/video, images and human speech (Russom, 2011). It is important to notice that data and information refer to separate but adjacent concepts. In fact they constitute the base components of both the Knowledge Pyramid (Song & Zhu, 2015) and the Data-Information-Knowledge-Wisdom hierarchy (Rowley, 2007). Within this study, we use the word data when referring to a raw collection of values, normally generated through recording of events, while information indicates the next level of contextualization and structure added to data with the purpose of enabling human cognition.
- **Technology**: The development of increasingly cheaper and more powerful technologies for storing, transmitting and processing data is one of the fundamental enablers of the rising of Big Data. Storing capacity of integrated circuits has grown exponentially over the last 50 years, as the density of transistors has nearly doubled every 24 months, following Moore's law (Moore, 2006). Processing data has become faster and cheaper thanks to the evolution of distributed computing and the availability of faster networks. For instance, a popular technology connected with Big Data today is Hadoop, an open source framework that lets clusters of dispersed machines co-operate in order to achieve higher performance through parallel computing (Davenport & Dyché, 2013). Another feature of Big Data Technology is the emergence of cloud

Please cite this article as: A. De Mauro et al., Human resources for Big Data professions: A systematic classification of job roles and required skill sets, Information Processing and Management (2017), http://dx.doi.org/10.1016/j.ipm.2017.05.004