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

De Mauro Andrea, Greco Marco, Grimaldi Michele, Ritala Paavo



**DOI:** 10.1016/j.ipm.2017.05.004



**Please cite the publication as follows:**

De Mauro, A., Greco, M., Grimaldi, M., Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing and Management. Volume 54, Issue 5. Pp. 207-817. DOI: 10.1016/j.ipm.2017.05.004

# Human Resources for Big Data Professions: A systematic Classification of Job Roles and Required Skill Sets

Andrea De Mauro *

Department of Enterprise Engineering
University of Rome Tor Vergata
Via del Politecnico, 1, 00133, Rome, Italy
E-mail: andrea.de.mauro@uniroma2.it
* *Corresponding author*

Marco Greco

Department of Civil and Mechanical Engineering
University of Cassino and the Southern Lazio
Via G. Di Biasio, 43, 03043, Cassino (FR), Italy
E-mail: m.greco@unicas.it

Michele Grimaldi

Department of Civil and Mechanical Engineering
University of Cassino and the Southern Lazio
Via G. Di Biasio, 43, 03043, Cassino (FR), Italy
E-mail: m.grimaldi@unicas.it

Paavo Ritala

School of Business and Management
Lappeenranta University of Technology (LUT), Finland
E-mail: paavo.ritala@lut.fi

**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.

**Keywords –** Big Data, Business Intelligence, Human Resources Management, Machine Learning, Topic Modelling.

## 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*

De Mauro, A., Greco, M., Grimaldi, M., Ritala, P. (2018) Human resources for Big Data professions: A systematic classification of job roles and required skill sets, Information Processing & Management, Vol. 54 Issue 5, pp. 807-817
doi:10.1016/j.ipm.2017.05.004

*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 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, taking the lead from (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 (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 computing, which allows companies to keep fixed costs at a minimum, thanks to the pay-as-you-go financial model of cloud-based services.
- **Methods**: The usage of Big Data entails the adoption of novel analytical methods for the transformation of big quantities of information into insights and, hence, value for the business. Furthermore, such data is often ill-structured, embedding diverse forms of coupling relationships whose modeling is critical, yet very challenging (Cao, 2015). The peculiar features of Big Data force practitioners to rethink their traditional data analysis toolkits and deepen the expertise in statistics and machine learning (Manyika et al., 2011). A partial list of the more recent analytical techniques we most frequently encounter in Big Data applications include: Cluster analysis, Genetic algorithms, Natural Language Processing, Speech and Image recognition, Neural Networks, Predictive modelling, Regression Models, Social Network Analysis, Sentiment Analysis, Signal Processing and Data Visualization (Chen, Chiang, & Storey, 2012; Manyika et al., 2011; Marr, 2015).
- **Impact**: The rise of Big Data has pervasively impacted a myriad of aspects of human life, ranging across science, economics, culture and society, in both positive and negative ways. Businesses are given the opportunity to create economic value through the analysis of Big Data. According to Davenport (2014), Big Data can drive companies through cost reduction, improvements in decision making and improvements in products and services. As a result, companies that have more aggressively shown their interest in Big Data tend to be more productive than their industry peers (Bughin, 2016). Big Data carries also widespread concerns of adverse impact on society, companies and individuals (Boyd & Crawford, 2012). The biggest concern is related to privacy: datasets carrying digital traces of a person's life can be used to uncover private details or even predict the future behaviour of individuals. This makes privacy a primary concern for those companies who want to establish a sustainable use of Big Data when interacting with consumers.

Information, Technology, Methods and Impact correspond to the most critical components of Big Data and are explicitly called out in the consensual definition which we reported in the introduction (De Mauro, Greco, & Grimaldi, 2016; van Altena et al., 2016). The emergence of novel approaches across each of these components has brought considerable challenges for human resources management within existing companies. The advent of new sources of data coupled with the renewal of methods and technologies used for business-impacting analytics require the development of new interdisciplinary competencies spanning from IT skills to business domain knowledge and communication skills (Chen et al., 2012). This poses a talent challenge for companies, seeking to upgrade their human capital, and for educators, who need to prepare the future generation of Big Data professionals and Analytics-savvy managers.

## 2.2 Myths and truths on Big Data jobs

The surge in popularity of the term 'Big Data' has been shortly followed by the establishment of another popular expression, strongly connected with the previous one, which had an exponential increase in popularity as of late 2012: 'Data Science'. Figure 1 shows the steep increase in web searches including the terms "Data Science" and "Data Scientists", obtained through Google Trends tool.

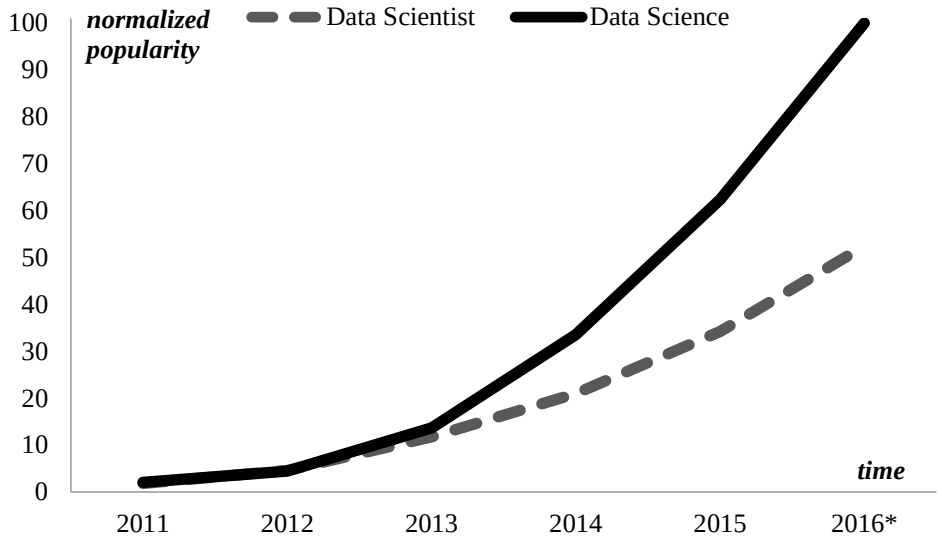


*Figure 1: Popularity of 'Data Science' and 'Data Scientist' among web users between 2007 and 2016 (*: 2016 corresponds to average popularity through the month of October '16 only). Values are proportional to the number of queries run by Google Search users, normalized to the highest point in the chart. Source: Google Trends.*

Differently from more traditional disciplines of science, the boundaries of this field have not been formally clarified from an academic point of view. Provost and Fawcett suggest two reasons behind the

lack of clarity around the concept of data science (Provost & Fawcett, 2013). Firstly, data science is strongly associated and confused with the concepts of Big Data and Data-Driven Decision-making, DDD (Brynjolfsson, Hitt, & Kim, 2011). Secondly, this scientific discipline is still at the stage of being more practical and experimental versus theoretical and methodological. At this point of the discipline development, there is a natural tendency for people to confuse the definition of the field with the description of what the practitioners of the field - data scientists, in this case - do.

Literature provides multiple, partially concurring descriptions of the characters Data Scientists should display. We notice that these items could be conceptually split into two groups. The first comprises the abilities which concur to the transformation of raw data into insights, with an emphasis on technology and methods: this group includes more technical skills, such as the ability to use Big Data tools, and proficiency in statistics and machine learning methods. The focus of this first group is essentially towards data and systems and requires "harder skills" and specific data and analytics knowledge. The second group consists of the capacity to transform insights into organizational value creation, involving the right people and core business processes in the organization. This group includes soft skills, mainly in the area of communication, general management techniques and knowledge of a specific business domain. Table I includes a non-exhaustive list of the many 'traits' a data scientist may exhibit.

| Focus | Data Scientist Traits | Source |
|---|---|---|
| Data and Technology | Big Data Tools Expert | (Miller, 2014; Provost & Fawcett, 2013; Song & Zhu, 2015) |
| | Coder | (Davenport & Patil, 2012) |
| | Statistician/Quantitative Analyst | (Davenport, 2014; Provost & Fawcett, 2013) |
| | Researcher | (Davenport, 2014) |
| | Data Hacker | (Conway, 2010; Davenport & Patil, 2012) |
| | Auditor | (Mayer-Schönberger & Cukier, 2013) |
| | Data Ethical Manager | (Miller, 2014) |
| | Data Manager and Strategist | (Miller, 2014; Song & Zhu, 2015; Wixom et al., 2014) |
| | Visualization Expert | (Davenport & Patil, 2012; Provost & Fawcett, 2013) |
| Business | Communicator | (Chen et al., 2012; Davenport & Patil, 2012; Song & Zhu, 2015; Wixom et al., 2014) |
| | Project Manager | (Song & Zhu, 2015) |
| | Business Expert/Advisor | (Chen et al., 2012; Davenport, 2014; McAfee & Brynjolfsson, 2012) |

*Table I: Traits displayed by Data Scientists, as encountered in existing literature.*

In our opinion the blurred understanding around the concepts of data science and data scientist is having two adverse consequences: the first one is the creation of the misleading myth around the role of data scientist, which is seen as the single most important actor in the process of creating economic value

through the usage of data. The second consequence, linked with the previous, is the disregard versus the other fundamental players concurring to a mature exploitation of data in a firm. As also noticed by Miller (2014), the sole role of Data Scientist is not accounting for the full realm of talent requirements companies are experiencing within the area of Big Data Analytics. The fundamental heterogeneity of the many facets displayed in Table I is an indication of the confusion generated over this concept, which brought many roles and skills to be indiscriminately associated under the same umbrella term of 'Data Scientist'. To provide more understanding in this regard, the current study aims to provide clarity on the job roles and skills companies need to develop and retain in order to fully reap the benefit of Big Data.

# 3. Methodology

In order to produce a data-based classification of Big Data job families and skill sets, we have designed a 4-step methodology by combining a series of existing analytical practices. First, we have downloaded a substantial amount of related online job posts, by means of web scraping techniques. Second, we have analyzed the words occurring in the job post titles in order to categorize them within a number of job families through expert judgment. Third, we have identified relevant skill sets by applying a topic modeling algorithm on the content of the job posts. Fourth and final step was to assess the relative importance of skill sets within job families by analyzing the average degree of presence of each skill set within job posts of each family. In the following sections, we describe every step in more detail.

## 3.1 Web Scraping

The World Wide Web contains a vast amount of information in various forms and levels of structure. In order to leverage such information, a computer program or an automated script (i.e. web crawler) can systematically retrieve webpages and store them in a central location (Kobayashi & Takeda, 2000). Web scraping consists in looking for specific data elements of interest from a series of semi-structured web pages, extracting them through crawlers and storing them into more structured data sets (Vargiu & Urru, 2012). In order to retrieve a large number of online job posts related to Big Data we have used a web scraping tool, as previously done by Capiluppi and Baravalle (2010), to retrieve the title and the job description of every job post, create a database on the basis of the subsequent analytical steps in this study.

There is a number of commercial and open-source web crawlers and web scraping tools available. However, a thorough survey of the existing tools goes beyond the scope of this paper (Girardi, Ricca, & Tonella, 2006; Özacar, 2016). For the sake of this study, we have decided to use Portia[1], a visual web

[1] Portia is available at: http://scrapinghub.com/portia.

scraper that can be configured through a web-based guided procedure. We tested Portia over a number of online job websites in order to assess the feasibility of retrieving a relatively clean list of relevant job post titles and description, considering the complexity of the web pages structure and the efficacy of the search functionality.

We have compiled a list of online job websites, annotating their characteristics in terms of quantity of Big Data-related job posts, geographic scope and overall feasibility of web scraping. Table II shows the results of our assessment across job post websites. The last column reports our qualitative indication of web scraping feasibility: this depends on the level of standardization of job post pages HTML structure and is linked with the quality of scraping results obtained through Portia. After assessing these features we have selected Dice.com as a source of our study as it is the website that presented the best relative assessment across the parameters mentioned above.

| Website | # of relevant posts | Geographic scope | Web scraping feasibility |
|---|---|---|---|
| Dice | 1.000+ | USA | ● ● ● |
| Careerbuilder | 1.000+ | Global | ● ● |
| Glassdoor | 100+ | Global | ● ● ● |
| Experteer | 1.000+ | Europe | ● ● |
| Infojobs | 100+ | Italy, Spain, Brazil | ● ● ● |
| Indeed | 10.000+ | Global | ● |
| Monster | 1.000+ | Global | ● ● |
| Simply hired | 10.000+ | Global | ● |
| Jobrapido | 1-100 | Italy | ● ● |
| Linkup | 1.000+ | USA | ● |

*Notes: The number of dots in the last column indicate how feasible it was to scrape from each website (● = it is not possible to download quality results; ● ● = some elements of text can be downloaded with varying quality; ● ● ● =nearly all posts can be successfully retrieved).*

*Table II: Website selection matrix.*

We have let Portia run a daily web scraping session overall job posts carrying the exact phrase 'Big Data' within title or description for around 2 continuous months during the fall of 2015. After removing all duplicate and incomplete entries we were left with a data set of 2.786 job posts, which we have used as an input for the analysis of job families and skills.

## 3.2 Identification of Job Families

After retrieving the full list of job posts we have applied basic text mining techniques and expert judgment in order to identify the essential Big Data job families. In order to do so, we have calculated all possible couples of adjacent words (bigrams) appearing in the job titles and we have sorted them by decreasing number of occurrences. Then we have collectively reviewed the list of the most frequent bigrams: by considering the type of roles we have found in literature (see section 2.2) we were able to recognize 4 essential groups of job roles:

1. Business Analysts (Business-facing analysts, project/program managers, Business advisors)
2. Data Scientists (Quantitative analysts, Statisticians, Modelers)
3. Developers (BI coders, Machine learning implementers)
4. System Managers (Architects, Infrastructure Admins)

For each job role, we have identified the most frequent bigrams: Figure 2 shows the word cloud with the top 50 words recurring in the job titles we have analyzed while Table III reports the top bigrams falling within each identified job family. On the whole, we were able to categorize 69% of the total list of downloaded job posts into non-ambiguous families.


*Figure 2: Word cloud showing the top 50 words recurring in the Job Title. The font size of each word is proportional to the number of occurrences of each word.*

| Business Analyst | Data Scientist | Developer | Engineer |
|---|---|---|---|
| Project Manager | Data Engineer | Software Engineer | Data Architect |
| Business Analyst | Data Scientist | Java Developer | DevOps Engineer |
| Product Manager | Data Analyst | Hadoop Developer | Solution Architect |
| Program Manager | Data Consultant | Software Developer | Systems Engineer |

*Table III: Top occurring bigrams in job titles, grouped by job family.*

### 3.3 Identification of Skill sets

The objective of the third phase of the process was to cluster skills within homogeneous groups, that we call skill sets. In this context, homogeneity refers to the reasonable assumption that skills belonging to the same skill set (like 'risk management' and 'planning' within the skill set of 'Project Management') are more likely to appear together in the same job descriptions. We can draw an immediate analogy with topics within text documents: each document contains homogeneous groups of keywords that characterize the topics dealt by the document. It is important to notice that multiple skill sets, in different proportion, can be required by one single job role. Hence, a traditional clustering technique based on mixture model (such as k-means) would not suffice to represent the complex set of competency requirements included in job posts. Instead, we needed to rely on mixed-membership models (Airoldi, Blei, Fienberg, & Xing, 2008) where the assumption that a unit belongs to a single

cluster is violated (Airoldi, Blei, Erosheva, & Fienberg, 2014). For the sake of identifying skill sets within job posts, we decided to adopt the mixed-membership model Latent Dirichlet Allocation, LDA (Blei, 2012), which has proven to work effectively at analyzing user-generated content like job posts (Ma, Zhang, Liu, Li, & Yuan, 2016).

LDA uses Bayesian Estimation Techniques in order to infer a vector representing the degree of membership (topic proportion) of each element (document) to each group (topic). By applying LDA, each topic can be seen as a distribution over the dictionary of words included in the corpus of the documents under study. The list of the most common words within a topic (keywords) can be used by an expert to deduce a meaningful description of the topic. For the sake of the current study, we have applied LDA on the description of the job posts: the 'keywords' referred to 'job skills' while the concept of 'topic' was substituted by the one of 'skill sets'.

As suggested by Moro et al. (2014), in order to keep the scope within a manageable list of skills, we have defined a dictionary that encompasses the more common terms which could be unambiguously linked to relevant skills within the domain of Big Data Analytics. We have again used the R package 'tm' (Feinerer, Hornik, & Meyer, 2008) in order to run LDA on job descriptions and retrieve the skills required in Big Data jobs.

The inputs to LDA are the input documents to be analyzed (in our case the job descriptions downloaded from dice.com) and the number of topics $k$ to be identified. As suggested by Chang *et al.* (2009), and confirmed by Blei (2012), we can select $k$ by applying human evaluation among alternative values, so that the interpretation of the machine-generated model results as meaningful as possible for a human. The authors have collectively evaluated multiple outputs of LDA with $k$ ranging from 2 to 30 and have consensually agreed that the most significant set of topics was reached with $k$=9.

Table IV shows the 9 skill sets we have identified through LDA: for each of those we have provided a title and the list of the top 20 keywords associated within each skill set.

| **Skill set #** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| **Title** | **Cloud** | **Coding** | **Database Management** | **Architecture** | **Project Management** | **Systems Management** | **Distributed Computing** | **Analytics** | **Business Impact** |
| **Top keywords** | Product | Software | Database | Design | Management | Systems | Hadoop | Analytics | Business |
| | Engineering | Applications | SQL | Technology | Project | Testing | Java | Science | Services |
| | Cloud | Engineering | Tools | Technical | Business | Security | Platform | Problems | Solutions |
| | Services | Web | Modeling | Solutions | Planning | Support | Distributed | Computer | Consulting |
| | Network | Code | Server | Architecture | Communication | Information | NoSQL | Learning | Technology |
| | Infrastructure | Applications | Oracle | Applications | Responsibilities | Tools | Hive | Analysis | Information |
| | Software | Java | Support | Architecture | Process | Monitoring | Source | Programming | Market |
| | Technology | Technology | Etl | Lead | Analyst | Programming | Python | Solving | Delivery |
| | Platform | Agile | Reporting | Responsibilities | Lead | Responsibilities | Spark | Applications | Financial |
| | Computer | Javascript | Process | Leadership | Change | Software | Hbase | Statistical | Delivering |
| | Deployment | Federal | Intelligence | Methodologies | Execution | Scripting | Scale | Modeling | Management |
| | Storage | Mobile | Design | Strategic | Risk | Technical | Pig | Languages | Sales |
| | Cisco | APIs | Business | IBM | Reporting | Document | Scripting | Algorithms | Strategic |
| | Management | Compliance | Queries | Document | Support | Linux | Cluster | Machine | Lead |
| | Responsibilities | Spring | Document | Volume | Agile | Applications | Process | Techniques | Execution |
| | Connected | Stack | Microsoft | Network | Excellent | Troubleshooting | Cassandra | Sets | Solving |
| | Virtualization | Rubiks | Communication | SDLC | Objectives | Problems | MapReduce | Excellent | Communication |
| | Scale | Scalable | Warehouse | Analytics | Track | Communication | Scalable | Predictive | Support |
| | Delivery | Network | Source | Deep | Document | Process | Linux | Product | Objectives |
| | Internet | Html | Analyze | Deployment | Programming | Debugging | Deployment | Scientist | Process |

*Table IV: The 20 most popular keywords referring to skills, grouped by skill sets, as per LDA output. The title in bold is a human interpretation of the generic focus of each skill set.*

### 3.4 Mapping of Skill Sets by Job Family

The objective of the fourth and last step of our analytical process was to characterize each job family recognized within phase 2 with a mix of relevant skill sets, identified in phase 3.

The presence (or topic proportion, in the context of topic modeling) is a measure of the extent at which each skill set is represented within each job post description. This measure is an output of LDA and is stored as a "presence matrix" where the rows are the job posts and the 9 columns correspond to the skill sets. By analyzing the presence of each skill set within each job description we have inferred the degree of centrality of each skill set in each job family. In order to do so, we have first grouped the presence factors by job family, using the classification obtained through phase 2 and calculated the average presence of each skill set within every job family. The resulting matrix $\mathbf{C}$ showed the average level of centrality of a skill set within every job family. We have then normalized matrix $\mathbf{C}$ by dividing every column by its average values, obtaining matrix $\hat{\mathbf{C}}$, whose elements $\hat{c}_{i,j}$ can be used to assess the centrality of each specific skill $j$ in a job family $i$ by means of the following relation:

- $\hat{c}_{i,j} > 1$: skill $j$ is typically relevant within job family $i$;
- $\hat{c}_{i,j} < 1$: skill $j$ is not typically relevant within job family $i$.

Table V shows a categorical assessment of relevancy of skills by job family.

| | Cloud | Coding | Database Management | Architecture | Project Management | Systems Management | Distributed Computing | Analytics | Business Impact |
|---|---|---|---|---|---|---|---|---|---|
| **Business Analyst** | | | • | | • • • | • | | • | • • • |
| **Data Scientist** | | | • • | | • • | | | • • • | • • |
| **Developer** | • • | • • • | • | • | | • • | • • | • | • |
| **Engineer** | • • | | • | • • • | • | • • | • • | | • |

*Notes: the number of dots in each cell indicate the relevancy of a Big Data Skill set within a Big Data job family and is based on the matrix* $\hat{\mathbf{C}}$ *(*$\hat{c}_{i,j} < 0.85$ *→ no dots, (*$\hat{c}_{i,j} \in [0.85, 1]$ *→ 1 dot,* $\hat{c}_{i,j} \in [1, 1.15]$ *→ 2 dots,* $\hat{c}_{i,j} > 1.15$ *→ 3 dots).*

*Table V: Big Data Job Families vs. Skill sets:*

# 4. Discussion

In this section, we provide an identikit of the job roles belonging to each Big Data job family. This description was obtained primarily by leveraging the family vs. skill set assessment which we obtained

through the process described above, and that we convey graphically through the alluvial diagram reported in Figure 3. We have also read a sample of job posts for each family and extracted some recurring features which helped to characterize the family. In the following section, some snippets of text (within quotation marks and in italic) are extracted from the downloaded job posts in order to illustrate actual examples the type of responsibilities related to each job family.

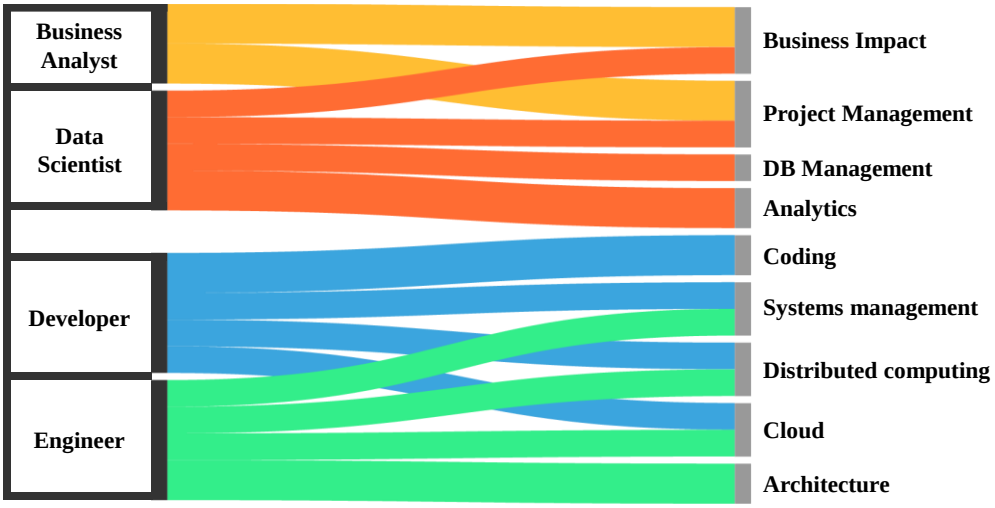


*Figure 3: Alluvial Diagram of Big Data Job Families vs. Big Data Skill sets. The width of each stream field is proportional to* $\hat{c}_{i,j}$.

## 4.1 Job Family 1: Business Analyst

The role of a Business analyst focuses on transforming relevant insights into actual business impact and includes elements of organizational effectiveness. Job posts within this family mention responsibilities in the area of analytical business advisory *("drive decision making through analytics, influence sales and marketing strategies, provide analytical support to business initiatives, report strategic insights to partners"*) and project management (*"analyze and document business needs, communicate progress and results effectively, bring to life recommended actions"*). Our research shows that the primary skills for a Business Analyst are in the domain of project management and business impact: we have used this last term to identify a mix of industry-specific skills and broader management competencies, such as effective communication, business process transformation and financial acumen. Business Analysts are the bridge between business decision makers and more technical roles: as a consequence, they also have a practical understanding of analytical methods and related technology which they relate to mainly as users.

## 4.2 Job Family 2: Data Scientist

The focus for a data scientist is on data itself and on the analytical methods for the transformation of data into insights. Job roles in this family include the responsibility to "*identify patterns, apply context*

*and intelligence, extract relevant information hidden in the large volumes of data, design and implement data models and statistical methods, integrate research and best practices into problem avoidance and continuous improvement"*. Posts usually mention specific analytical techniques ("*classification, collaborative filtering, association rules, neural networks, heuristic approaches*"), scripting or programming languages ("*Python, SQL, Java, Ruby*") and statistical platforms ("*R, SAS, Matlab*") deemed to be critical for the advertised position. According to our analysis, Data scientists' main skill set is definitely analytics, as they know and leverage Big Data Methods like anyone else in their company. Data Scientists need also to understand the business context in which they operate and use project management techniques in order to interact effectively with the rest of the organization. They should also be confident in accessing corporate data warehouses and can write scripts for querying databases.

### 4.3 Job Family 3: Big Data Developer

The main objective of Big Data Developers is to design, develop and modify data-reliant application software. Job descriptions within this family mention that candidates will *"develop dashboards and data solutions, design, build, and deliver new reports, prototype working proof of concepts for multi-threaded, multi-server applications, integrate third party applications through Application Programming Interfaces"*. Posts also refer to responsibilities over the application lifecycle of the analytical product, which include *"design, development and implementation of automation innovations, development of automated testing scripts, on-going advanced application support").* Their primary skill is undoubtedly coding, but they also need a solid expertise in systems management, cloud computing and distributed technologies. Big Data developers also require a basic understanding of database management, corporate data architecture, and need to know how analytics are used in the context of their company.

### 4.4 Job Family 4: Big Data Engineer

Big Data Engineers focus on building and maintaining the full technology infrastructure which enables storage and processing of Big Data. Roles in this family are responsible to*: "manage the enterprise analytics server platform, support all processes to load and manage the analytics data store and integrate new data sources, ensure capacity, backups, failover, and disaster recovery processes are in place, deploy custom cloud-based applications, scale backend data storage platforms."* Job posts usually include specific mentions of the technologies adopted in the company's Big Data stack and might include (in no specific order) "*Hadoop, Cassandra, MongoDB, MySQL, Hana, Ceph, GlusterFS, Azure, Amazon Web Services*". The primary skill set for Big Data Engineers relates to data architecture and includes the competencies needed to construct and manage the corporate Big Data ecosystem in a sustainable manner. This includes the ability to take care of the variety of complexities inherent to

systems management (spanning from information security to performance monitoring), cloud computing and distributed processing. Job posts also refer to the ability to interact with databases, adopt project management processes and have a general understanding of how data can support the company's strategy.

# 5. Conclusions and Managerial Implications

Data scientists have been put under the spotlight as the - supposedly - protagonists of the Big Data revolution in companies (Davenport & Patil, 2012). Firms need to get the right analytical skills and expertise added to their human capital but this goes well beyond acquiring data scientists alone. Managers still question themselves on which new talent they need and on how to upgrade the skills of their current human resources (Davenport, 2014; McAfee & Brynjolfsson, 2012). With the present study, we have provided structure and clarity to the multifaceted landscape of Big Data-related human resource needs, by offering a systematized nomenclature and characterization of job roles and skills. This contributes to the literature by enabling a coherent framework upon which to build future investigations, as desired my multiple researchers (Miller, 2014; Song & Zhu, 2015)

We have assembled a semi-automated analytical process, based on web scraping, expert judgment, text mining and topic modelling techniques in order to systemically review the current job offers related to Big Data, using more than 2.700 job descriptions posted online. Our findings confirm the ideas of Miller (2014), who suggested that Data Scientists and their deep expertise on Analytical methods are far from being sufficient in granting companies a real competitive advantage. The evidence from our analysis suggests that there are 4 different job families related to Big Data, which are: Business Analysts, Data Scientists, Big Data Developers and Big Data Engineers. We have characterized each of them with a data-based assessment of the skill sets required by each role family and the required level of proficiency. We have built a 'Big Data Job Families vs. Skill sets matrix' (Table V) which can be used by business managers to structure their recruitment programs and functional career paths and also by universities for the sake of shaping their curricula and degree programs. The matrix also suggests a natural clustering of Big Data Job Families into two separate groups:

- *Technology-Enabler professionals*: Developers and Engineers have multiple overlaps in terms of skill sets and their role tend to be more technical-facing and focused on systems and applications;
- *Business-Impacting professionals*: Business Analysts and Data Scientists share multiple skill requirements and have a more business-oriented role, focusing on data analysis, in direct connection with economic impact and organizational value creation.

There results provide particularly useful insights for organizations and managers working in industries that are transformed by 'digital disruption'. For instance, functional managers can use our results to build more meaningful and structured job descriptions for hiring. Moreover, HR managers can design Big Data career and competency development frameworks in a way that is coherent with the most prominent business needs and industry trends. The results also provide useful guidance to educational institutions (such as universities and their masters' programmes) that aim to focus their efforts in developing skills and competences that are needed in the future. To this end, our results suggest that Data Scientist is not a profession which is homogeneous, but includes both 'hard' and 'soft' skills, as well as different connotations towards organizational processes, technologies, and value creation. Thus, educational programmes could be tailored accordingly, taking into account the specific industry needs.

Beside the above-described results, the present study contains an original methodological proposal that can be reapplied to bring clarity to other domains. In fact, the conjoint use of web scraping and machine learning techniques for classifying jobs and describing them in terms of skill requirements is innovative and can be reused in similar future studies focusing on any other professional field.

The current investigation is affected by a number of known limitations which provide stimulating opportunities for future research: firstly, the analysis of job posts was largely based on US-based positions and might not consider relevant trends generated in other geographical regions. Secondly, we provided clarity upon the features of job roles, each considered individually, while neither offering details on their mutual interactions, nor examining how the organization as a whole should be structured. Lastly, the characterization of the various Big Data skill sets does not provide a precise indication of the behaviors professionals should display: the addition of a proficiency assessment tool for those skills would enable companies to rate the maturity of their analytical workforce.

**Acknowledgements**

The authors would like to thank Giacomo Nobili for his valuable contribution.

# References

Airoldi, E. M., Blei, D. M., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2014). *Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL: CRC Press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9,* 1981–2014.

Blei, D. M. (2012). Introduction to Probabilistic Topic Models. *Communications of the ACM*, *55*, 77–

84.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*, 662–679.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*, 1–28.

Bughin, J. (2016). Big data, Big bang? *Journal of Big Data*, *3*, 2.

Cao, L. (2015). Coupling learning of complex interactions. *Information Processing & Management*, *51*, 167–186.

Capiluppi, A., & Baravalle, A. (2010). Matching demand and offer in on-line provision: A longitudinal study of monster.com. *Proceedings - 12th IEEE International Symposium on Web Systems Evolution, WSE 2010*, 13–21.

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288--296.

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*, 1165–1188.

Colombo, M. G., & Grilli, L. (2010). On growth drivers of high-tech start-ups: Exploring the role of founders' human capital and venture capital. *Journal of Business Venturing*, *25*, 610–626.

Conway, D. (2010). The Data Science Venn Diagram.

Davenport, T. H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press.

Davenport, T. H., & Dyché, J. (2013). *Big Data in Big Companies*. Portland, OR: International Institute for Analytics.

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job Of the 21st Century. *Harvard Business Review*, *90*, 70–76.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*, 122–135.

De Mauro, A., Greco, M., Grimaldi, M., & Nobili, G. (2016). Beyond Data Scientists: a Review of Big Data Skills and Job Families. In J. . Spender, G. Schiuma, & J. . Noenning (Eds.), *International Forum on Knowledge Asset Dynamics 2016* (pp. 1844–1857). Dresden.

Delgado-Verde, M., Martín-De Castro, G., & Amores-Salvadó, J. (2016). Intellectual capital and radical innovation: Exploring the quadratic effects in technology-based manufacturing firms. *Technovation*, *54*, 35–47.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of …*, *25*.

Gandomi, A., & Haider, M. (2014). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*, 137–144.

Girardi, C., Ricca, F., & Tonella, P. (2006). Web crawlers compared. *International Journal of Web Information Systems*, *2*, 85–94.

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34*, 135–174.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, *32*, 144–173.

Ma, B., Zhang, N., Liu, G., Li, L., & Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management*, *52*, 430–445.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. *McKinsey Global Institute*.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Wiley.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *90*, 61–67.

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills Gap. *Journal of Organization Design*, *3*, 26–30.

Moore, G. E. (2006). Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Newsletter*, *11*, 33–35.

Morales-Alonso, G., Pablo-Lerchundi, I., & Núez-Del-Río, M. C. (2016). Entrepreneurial intention of engineering students and associated influence of contextual factors. *International Journal of Social Psychology*. doi:10.1080/02134748.2015.1101314

Moro, S. M. C., Cortez, P. A. R., & Rita, P. M. R. F. (2014). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*, 1314–1324.

Özacar, T. (2016). A tool for producing structured interoperable data from product features on the web. *Information Systems*, *56*, 36–54.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Data Science and Big Data*, *1*, 51–59.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*, 163–180.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report*.

Saenz, J., Aramburu, N., Buenchea, M., Vanhala, M., & Ritala, P. (2017). How much does firm-specific intellectual capital vary? Cross-industry and cross-national comparison. *European Journal of International Management*, *11*, 129–152.

Siepel, J., Cowling, M., & Coad, A. (2017). Non-founder human capital and the long-run growth and survival of high-tech ventures. *Technovation*, *59*, 34–43.

Song, I.-Y., & Zhu, Y. (2015). Big data and data science: what should we teach? *Expert Systems*. doi:10.1111/exsy.12130

van Altena, A. J., Moerland, P. D., Zwinderman, A. H., & Olabarriaga, S. D. (2016). Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data*, *3*, 23.

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*, *2*, 44–54.

Ward, J. S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions, 2. Databases.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., … Turetken, O. (2014). The current state of business intelligence in academia: The arrival of big data. *Communications of the Association for Information Systems*, *34*, 1–13.

Ylijoki, O., & Porras, J. (2016). Perspectives to Definition of Big Data : A Mapping Study and Discussion. *Journal of Innovation Management*, *1*, 69–91.