ISSN 2411-9563 (Print)
ISSN 2312-8429 (Online)

European Journal of Social Science
Education and Research

September 2018
Vol 5 No 3

DOI: 10.2478/ejser-2018-0065

Organizational Design of Big Data and Analytics Teams

Lennart Hammerström

University of Kaposvár, Hungary

Abstract

Although many would argue that the most important factor for the success of a big data project is the process of analyzing the data, it is more important to staff, structure and organize the participants involved to ensure an efficient collaboration within the team and an effective use of the toolsets, the relevant applications and a customized flow of information. A main challenge of big data projects originates from the amount of people involved and that need to collaborate, the need for a higher and specific education, the defined approach to solve the analytical problem that is undefined in many cases, the data-set itself (structured or unstructured) and the required hard- and software (such as analysis-software or self-learning algorithms). Today there is neither an organizational framework nor overarching guidelines for the creation of a high-performance analytics team and its organizational integration available. This paper builds upon (a) the organizational design of a team for a big data project, (b) the relevant roles and competencies (such as programming or communication skills) of the members of the team and (c) the form in which they are connected and managed.

Keywords: Big Data and Analytics • Organizational Design • Roles and Competencies • Incorporation of teams

Introduction

Big data and analytics is not a vision, an idea or a concept only for very specific fields of application any more. It is driven within almost all industries and there are many success stories told within a wide range of different industries.

The possibilities to analyze big data are enhancing and what is possible today could hardly be imagined a decade ago. Also the toolset has developed dramatically (e. g. deep learning and neuronal networks). It provides new approaches to the researcher and enables the applicant to search through more and more data and to find patterns with automatisms.

In many cases the organization for big data projects focus on the application of tools and software, a powerful hardware to crunch the vast amount of data and sophisticated algorithms, applicable by specialists only.

But even more important, due to the complexity and the variableness of big data projects, is it to find the right setup of roles and competencies in the earliest project phase. It is also necessary to take social skills such as the ability to communicate and to share information into consideration.

For most of the technical problems coming along with analytics applications can be purchased; expertise, processing power and data storage can be hired or leased at external companies. But big data projects require often a higher education and a further training on specific fields of applications.

The organization design of the team and its integration into the companies' processes is often disregarded. Working out the organization design does not only have to take into consideration the companies structure but also the requirement of different roles and the internal and external collaboration.

Setting up a big data analytics team in the traditional way (e.g. within an own department with clear boundaries) is leading to accurate results but it limits the operating distance of the team and produces segregated applications. The team will not be interconnected exceeding for example a business unit or a certain type of departments (such as Research and Development).

To attain the goal of each big data project (or to exceed its expectations) the applications need to be connected and embedded into an overall digital manufacturing strategy; a flow of information and a common understanding of the problems has to be achieved and comprehensive expectations of the "success" of the projects has to be obtained.

This paper builds upon

The **organizational design** of a team for a big data project

The **relevant roles and competencies** (such as programming or communication skills) of the members of the team and

The form in which they are **connected and managed**.

Some further comments on big data and analytics are reasonable to capture the relevance of the organizational design for such type of projects.

Applying big data and analytics

There is no rigorous definition for Big Data. Dijcks states that it refers to three different types (Dijcks, 2013):

Traditional enterprise data

Machine- or sensor generated data and

Social data

Chen (Chen et al 2012) defines big data in the volume of data, starting at terabyte or petabyte, sometimes even at exabyte (Chen et al 2012). Only some years ago data-sets with some hundreds of thousand observations were considered to be 'big', today this is an averages sample size (Reimer et al. 2014) (Kübler et al. 2017).

Besides the volume (tera- or petabyte) there is velocity (the data-flow is too fast for analysis), variety (the range of data types) and variability (the data is unpredictable or erratic) of the data. The data derives from different sources, such as the environment.

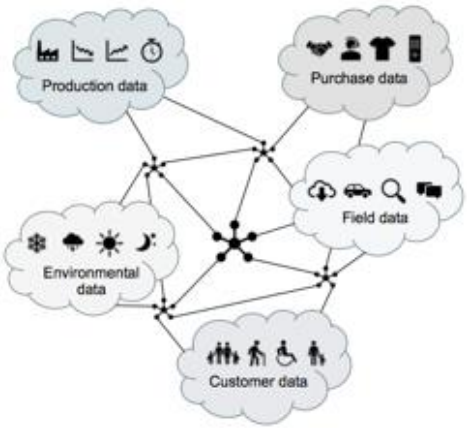


figure 1 - multiple sources of data

With the turn of the millennium the volume of data started skyrocketing, the available technology back then was not able to analyze the amount of data; up to the point that IT faced a data scalability crisis (Dijcks 2013). Due to further development

of the relevant technologies and following Moore's law, the management and the analysis of this tremendous amount of data became possible.

Currently most of the publics' attention is drawn to the influence of social data (e.g. the data scandal that shocked Facebook, triggered by Cambridge Analytica) and its impact on the market (Facebook shares fell 5% within the scandal, worth $80b (Monica, 2018).

The huge positive impact of the analysis of big data is often not that present in media and therefore unknown to the public. Internet giants like Google (e.g. Google Bigtable; designed to scale petabytes of data or Google Dremel; designed to run instant queries on multiple petabytes of data in seconds) (Scott, 2014) or Facebook (e. g. viability to predict highly sensitive personal attributes by analyzing a user's account and likes) (Monnappa, 2017) developed outstanding solutions to handle big data.

There is also a negative side of big data and analytics, for example the tendency to always ask for more data – just because we can (Håkonsson et al. 2016). This can slow down decision processes because more and more data are requested and can lead to the attitude to let the machine find and make management decisions. This side of the analysis of big data-sets is left unattended in this article.

For companies such as Google, Facebook and others, the collection of data has become an end in itself rather than a way of achieving other ancillary business goals (Nunan and Domenico, 2015).

This is a strong indicator that the way businesses compete, collaborate, and operate is changing. The competition is accelerating and encompasses the next-generation competition (Teece, et al. 2017); and big data is part of the driver for this acceleration.

Facebook is the social website with the largest number of users [more than 2 billion users per month in 2017 (YouTube: 1.5b, WhatsApp: 1.2b, WeChat: 889m) (Constine, 2017)]. In 2014 Facebook stored a volume of 300PB and had a daily incoming rate of 600TB (Vagata and Wilfong, 2014). This data is used to explore connections and relations between users (and even none-users that don't even have a Facebook profile), their needs and expectations and their behavior to improve for example the response-ratio for advertisement. To analyze a data-set with exabytes is exceptional and challenging, but it is possible (Chen et al. 2012).

Today Big Data and Analytics solutions can be applied to almost any industry. The high-impact areas of today are the e-commerce and the market intelligence, but also within the health industry and security and safety the capabilities of the analyses have been proven.

The data itself is not generating value, if not purchased as data to a third party. But with the application of analytical methodologies a value can be created. Big Data and Analytics can provide insights into hidden patterns and relations, help to come to the right conclusions and to improve the decision process.

The analysis itself is a conjunction of different methodologies and requires a different compound of methods, experts and applicants. The applied methods have to fit to the problem that shall be solved; otherwise the result is questionable. There are different methods available; such as logistic regression, decision trees, artificial neural networks, discriminant analysis, random forest and others (see the supplement). The use of one of these techniques requires expertise and experience to interpret the results in the right way and to come to the right conclusions.

The amount of data that is handled in manufacturing analytics projects is less than in social data analytics but reaches also giga- or terabytes of data for a single analysis.

Manufacturing companies recognized several years ago that the data they 'produced' on their shop floor has developed to a point that it has become an asset in its own. The analyst's task is to uncover patterns within the datasets and to discover new business facts that are currently unknown to the enterprise (Russom, 2011).

In most of the cases, a big data project will be started due to the tangible benefits that are expected to come with a successful project. The analysis can be utilized to support different goals of the enterprise. Most obvious is the improvement of established processes, for example the increase of efficiencies.

Though some of the benefits are underestimated in big data projects, intangible benefits (Hadjinicolaou et al. 2018) strike areas such as an improved relationship and/or an improved mutual trust with share- and stakeholders, organizational capabilities or legal/customer requirements.

Applying big data and advanced analytics, enterprises can also develop a better understanding of the current state of their businesses and also predict evolving aspects, such as future customer behavior.

Implementing a big data project requires not only to hire people with the right knowledge and the right mindset, but also to connect those with the experts in production, design and development, industrialization, process design and so on.

To that end, there is no such structure available today that facilitates the top management with the right tool set to staff such projects.

The article provides a framework to analyze the critical success factors, to rate the hard- and soft skills of the team members and a toolset to evaluate the concrete set up of the team. Weaknesses in the set-up can be identified right in the beginning and counteracting measures can be taken.

Origin of the Term 'Big Data'

It is not fully clear, who coined the phrase 'big data', but it is certain that its meaning was different when used by authors three decades ago.

The first who conjoined statistics, a storage medium (computer tapes) and the methodology to extract information from data was Charles Tilly in 1984.

"**Against these procedures, Stone lodges the objection that historical data are too unreliable, …, that the storage of evidence on computer tapes blocks** the **verification conclusions by other historians, …, that none of the big questions has actually yielded to the bludgeoning of the big-data people, …"** (Tilly 1984)

Other sources can be traced back till the 80s, such as an article from 1989 by Erik Larson writing for Harper's Magazine and showing the far-sightedness of the author (Larson 1989).

"The keepers of big data say they do it for the consumer's benefit. But data have a way of being used for purposes other than originally intended."

Also John Mashey, who was working at the beginning of the 90s for Silicon Graphics is credited to have invented the phrase. Mashey was using that term for a range of issues, looking for the simplest, shortest phrase to convey that the boundaries of computing keep advancing (Lohr 2013). There is no written paper or published article that can be used as evidence but there are former colleagues that testify that he was using the term.

A definition of Big Data

Mauro et al. published a substantial article regarding a consensual definition of big data (Mauro et al 2015. "What is big data? A consensual definition and a review of key research topics"). It is an extensive article that is recommended for further investigations and a more detailed research on the definition of big data.

The authors propose the following formal definition:

"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" (Mauro et al. 2015).

In October 2011, The Economist projected that the number of connected mobile devices would reach 10 billion in 2020 (The Economist 2011). In 2018 the number of connected devices already reached 23.14 billion devices, creating data non-stop (Statista.com 2018) demonstrating that the prediction accuracy is unsound.

Comments on Analytics

The analytics part is the key to success within a big data project from a mathematician's or engineer's point of view. It is the section where the extracted and prepared data is given to the data scientist and she has to apply mathematical models or work out an algorithm to apply machine-learning methods. To work out a sound model, the collaboration of experts from

different fields of expertise has to be applied. To work out and to apply an algorithm, experts of programming, data banks and mathematician have to contribute.

Two main directions have to be taken into consideration: statistics and machine learning.

The origin of those methodologies is a different one but machine learning and statistics share common methodologies such as regression analysis, resampling, classification and non-linear methods (Kübler et al. 2017).

Statistics

The methodology of statistics is applied in many different fields, such as physics, medicine, economy, astronomy, social science and the humanities.

The scientific foundation of the development of statistics can be seen in the work of academics such as Blaise Pascal[1] and Pierre Simon de Laplace[2] (and many others) working on probability calculations within the game of chance. The second point of origin was the 'description of the state', starting for example with the editing and aggregation of life- and mortality tables (Hudec & Neumann n. d.). Today there are different definitions eligible, such as:

Statistics is a science for the quantitative gathering and manageable preparation of mass occurring isolated phenomenon (Werth 1985)

or

Statistics is the methodology of learning based on empiricism (Hackl & Katzenbeisser 1996)

Statistics is classified in three subdomains (Moore 1992).

Descriptive statistic

Inductive statistic

Explorative statistic

Those three subdomains are all applied in the analysis of big data.

Wasserman claims that statistics in general is applied to low dimensional problems (Wasserman 2012). E.g. a statistical problem within chemical production is the correlation between the supplier of raw or pre-processed material and the yield of the final process. Problems such as regression, factor analysis clustering and discriminant analysis can be solved with multivariate statistics.

An important part of the analytics phase of a Big Data project is assisted from analytics software. There is proprietary software (such as Almo, GAUSS, Minitab, qs-STAT, SAS, SPSS, SsS) and non-proprietary software (such as PSPP, R, Statistiklabor) available and the selection of it is subject to the analytical problem.

Machine Learning

Machine learning is a sub-set of artificial intelligence (Marr 2016) and originates from computer sciences (Kübler et al. 2017).

In contrast to statistic, machine learning is more outcome-oriented and focuses on accurate prediction making. The data in computer science often originates from high dimensional problems and an undefined number of variables (Wassermann 2012).

[1] Pascal, Blaise (✶June 19, 1623 ✞August 19, 1662) was a French mathematician, physicist, inventor, writer and Catholic theologian (Adamson 1995).
[2] de Laplace, Pierre Simon (✶March 23, 1749 ✞March 5, 1827) was a French scholar who contributed to mathematics, statistics, physics and astronomy (Crosland 1967).

Machine learning algorithms use computational methods to "learn" information directly from the data-set without relying on a predetermined equation as a model, and the algorithms adaptively improve their performance as the number of samples available for learning increases (Soni 2017).

The machine learning system consists of three major parts, which are:

The model: the system that makes predictions or identifications

Parameters: the signals or factors used by the model to form its decisions

Learner: the system that adjusts the parameters – and in turn the model – by looking at differences in prediction versus actual outcome

Machine learning can be perceived as less restrictive and less formal than statistics (Wasserman 2012), but is facing other problems such as the required time (up to a week) to train a neural network (Milutinovic 2017).

Different technologies from statistics and computer science can be applied to analyze datasets. Manyika et al. provide a list (see supplement) of techniques applicable across a range of industries (Manyika et al. 2011).

Data Analytics Algorithms

There are several algorithms for the purpose of data mining and analytics. In 2006, the IEEE International Conference on Data Mining (ICDM) defined the 10 most influential data mining algorithms. Those are:

C4.5

k-Means

Support Vector Machines (SVM)

Apriori

Expectation Maximization (EM)

PageRank

AdaBoost

k-nearest neighbors (kNN)

Naïve Bayes

Classification and Regression Trees (CART)

Wu et al. (2007), Panson (2015), Quinlan (1986 and 2007) provide explanations regarding the algorithms and the area of their application. Especially the work of Wu et al. (2007) provides detailed explanations for the algorithms. For further readings on the algorithms, see the supplement.

Organizational Design

Organizational design is a systematic approach to aligning structures, processes, leadership, culture, people, practices, and metrics to enable organizations to achieve their mission and strategy the organization should be designed to fit the circumstances (Burton et al. 2018).

For the organizational design of a big data and analytics team, there is no such thing as the 'one best organizational design'. Before the team is set up, the following two main questions must be answered.

The first question is about the integration of the team within the company. How shall the lines of command be organized, what is the reporting structure, who is setting the targets?

The second question is about the staffing of the team. Before starting any hiring process, the corporation must define what type of problems need to be solved, how the required roles can be formed, what the strategic target of the big data team is and what competencies are required.

Integration

There are different factors that have an influence on the incorporation, such as:

Form of the organization

Functional

Divisional

Matrix

Channel of communication

Hierarchical

Network organization

Virtual organization

Spiral organization

The industry the company is operating in (e.g. manufacturing trade, information technology, commercial enterprise, chemical and process industry, textile industry, public services, energy provider, …)

Principle of structure

Origin of position (object oriented, execution, region)

Principle of management (unity of command, multiple-line system)

Decision making authority (centralized, de-centralized)

The size of the company (SME vs. large scale enterprise)

All topics are relevant, but the topics #1 and #2 are outstanding. Those are the ones that have to be considered first of all, the other ones have minor influence. The form of the organization and the channel of communication have to be defined upfront and have to be established to avoid losses in performance when the team shall start to work on the big data projects.

Big data and analytics requires a short chain of command, where information is flowing fast and without losses, the team has to have a company wide access to the data storage and the loggers to really get a full picture of the data and to ensure that all relevant information is taken into consideration.

The incorporation therefore has to enable the team members to get in contact with the relevant stakeholders of a company quickly and without obstructions to understand the requirements right from the beginning and to understand the scope of action as early as possible.

Roles and Competencies

The success of a project is in most cases measured by what is called the iron triangle: scope, deadline and cost. Defining success based on this three criteria is not easy and depends on the perspective of the share- and stakeholders and it is intensively influenced from the type of the project, the external and internal conditions, the prestige it obtains and so on.

Conventional projects can be organized, staffed and executed in a structured manner, especially if these projects are reiterating and even more when they are perseverating.

Big Data and Analytics projects have to be set up in a structured way as normal projects, too. But they cannot be treated in the same way; the complexity of these projects can hardly be handled by the standard project management techniques. The reason for this is:

The required competencies are very specific, the curriculums at the universities offer courses in this specific field of application since a little more than a decade and only as part of higher education (Debortoli et al. 2013)

As it was stated before, the interfaces within a project are complex; even with distinguished experts in every single phase of the project it can fail when information is transmitted incorrectly or is misunderstood or misinterpreted (Donald et al. 2013)

The interpretation of the analytics phase, its recommendation for the application and the transfer into the real life application requires a high level of awareness

With an increase of complexity, it is becoming necessary to invest more time for organizing and staffing the project in the right manner directly from the beginning.

As the integration of the team depends mainly on the current structure of the company and on the channels of communication, the links for the team and especially the manager of the project become important. Figure 2 shows a project manager within a matrix organization and the different communication channels she must serve.

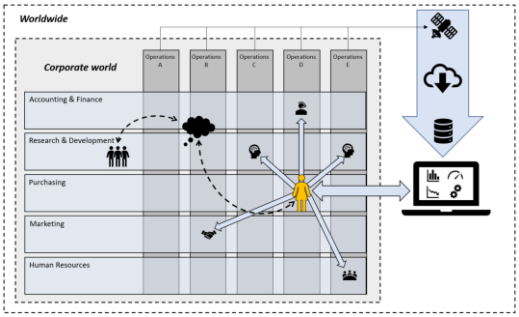


Figure 2 - Multiple levels of communication

The team must be able to share knowledge, share ideas and to work together to find the right solution for a certain problem. The interaction between the different areas is part of the success of innovative problem solution (Ebert, et al. 2017).

Required competencies

The literature regarding high-performance teams within the field of Big Data and Analytics is limited.

The research focuses either on the competencies (with a focus on analytics, less on social skills), or the management of the project, or the hardware (e.g. the drive), or the software (e.g. analytics software).

There is very little research available how to access the competencies, how to arrange a team with respect to their skills and the Big Data problem. The factors for the success of a Big Data project are not set into a concrete order to achieve the best out of the available resources.

Most authors share the opinion that the next-generation competition has turned the attention to superior competences and dynamic capabilities. This ability is necessary to coordinate the tangible and intangible recourses (Teece, et al. 2017).

Competencies that have been the key to success in one project can be futile in other ones. Therefore, it is necessary to evaluate the required competencies based on the specific project for each domain and each sub-domain before the start of a project (f*igure 3*).

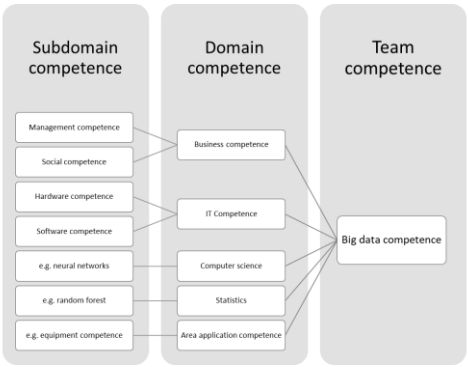


Figure 3 - Different domains of competences

An example of how demanding the tasks and how broad the range of the applications can become is provided in the supplement matrix *statistics and computer science technologies*; this matrix provides insights just for the analytical part of a Big Data project.

Required Roles

The contribution of the analytics toolset, the speed of the hardware and the latest version of the software that is used to process the data is too often the focus of attention. As a matter of course, there will hardly be a big data project without the support of hard- and software, no question about that.

However, without the right expertise, well-organized coordination of the project, and sound communication, all big data projects will fail or stay far beyond the expectations. Research has shown that "inadequate staffing and skills are the leading barriers to Big Data Analytics" (Russom 2011).

In addition to technical system implementation, significant business or domain knowledge as well as effective communication skills are needed for the successful completion of such BD&A projects (Chen et al. 2012).

To deploy and to implement a big data analytics team has a whole host of requirements and implications (Geissbauer et al. 2017). It is not only about hiring and making up the team, it is also about the team structure, the expertise of its members and the right composition of the team (Debortoli 2014).

Based on the content of the matrix in the supplement (roles within a Big Data team) the following roles within a Big Data team are proposed to be set up for a Big Data project:

Data analytics expert

Data scientist

Systems architect

Data software expert

Operations expert

Data project manager

The roles recommended cover the required competencies and the interfaces between the different scopes of application.

Each role has to cover a certain set of competencies to fulfill the tasks of a Big Data project.

Data analytics expert

The data analytics expert is the one who has to apply the statistics and computer science methodology to the data-set. It his her who has to find the patterns in the data and to point out where the team has to focus on and who has to make recommendations for the next steps.

Data scientist

Collecting, preparing and customizing the data from systems (like ERP or MES; databanks or data storage systems) are the task of the data scientist. She has to develop access to the relevant data, extract and provide it to the data analytics expert.

Systems architect

The systems architect creates an environment that is necessary to get access to the data, according to the requirements of the business. If the data, for example, is stored in a business intelligence system and the analysis requires a retrospective analysis, the architecture has to be set up in a certain manner. If the data has to be analyzed in real-time, the architecture has different requirements and will look different. The systems architect has to have a broad knowledge of IT systems, its structure and the advantages and disadvantages of different architectures.

Data software expert

The data software expert creates and establishes the computer system to carry out the Big Data operations. It can be considered as the 'backbone' of the analytics. It consists of a set of programs and procedures (some of them open source) that need to be modified for the application. It provides the ability to store and process huge amounts of data, the computing power (the more nodes, the more power) to the data software expert. It can be scaled according to the requirements.

Operations expert

Most big data projects zoom in on the analytics part and neglect the interfaces, the complexness of the 'real life' applications, or struggle to find the right interpretation what the results of the analysis imply.

This can result in overlooking or ignoring significant findings and thus the project falling short.

The team member who can provide the essential insights into the processes of the application, who can manage the tasks on the interconnection points and can conceptualize the projects broader framework, is compulsory.

The team member that can incur such liability is the operations expert.

Data project manager

One of the presuppositions of this article is that a big data project can only become an outstanding success if the required hard and soft facts are available and also managed in matters of the requirements of big data projects. This demands a specialized big data project manager.

The data project manager cannot be considered as a standard project manager and cannot be assigned from any other type of project to a big data project. Besides the iron triangle (scope, deadline and cost), she simultaneously has to take care of an intensive management of the interfaces; she has to invest much of her time in the management of the interfaces and has to coach the different experts that run the analytics part. Overall, she has to be well versed in cognitive and behavioral sciences and must be able to manage a team of highly educated experts.

Management of Interfaces

Treating a big data project in the same way as a conventional project can establish right at the beginning the reason for its fail at a later point in time. One of the key factors that require the project managers' attentiveness is the management of the interfaces.

The interfaces of a big data project cannot be mapped out in neat fashion with pre- or well-defined points of transfer. Deploying the analytical tools, the hardware that provides the required analytics performance and setting up the right systems architecture for a big data project are already a tough challenge, even with the right people in the team and an adequate budget at hand.

Liaising all of those different areas and experts is in the hands of the data project manager; but the team needs to be sensitized to the specific challenges and requirements of such projects and the delicate management of the interfaces.

Besides the standard tool set of a project manager the following tools can be advised to improve the management of the interfaces:

Full time assignment of team members to the project so they can focus

Setting up the team as kind of a 'virtual enterprise' with collaboration and communication tools (Cordova et al. 2013)

Combining the virtual tools with an agile approach for the project management to ensure that the team can maneuver fast and is not slowed down by external factors (Baker 2017; Contractor et al. 2006)

Access to external support; such as specialists/consultants from outside the company or senior experts within the enterprise when unexpected trials and tribulations come up

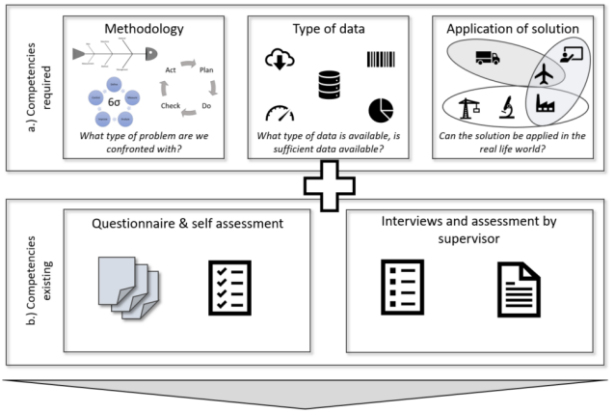
This kind of a team set up requires a high self-employment of the team members and a high degree of autonomy. Team members must possess a personality that supports this autonomous working style and an entrepreneurial spirit.

Besides the personality, the team must have the possibility to engage in so-called deep work and to focus on the problem and its potential solution. Knowledge workers need such distraction free time periods to finalize tasks that require extended periods of concentrated work (Ebert et al. 2017).

Assessment of the competencies

Before undertaking a big data project, it is recommended to (*a*) define the required competencies for this specific project and (*b*) to determine the existing competencies of potential team members. This needs to be done to match the requirements and the available competencies. If the requirements of a certain project cannot be fulfilled, it shouldn't be started in the first place.

Defining topic (a) can be done with an analysis of the existing situation. This requires (I) an evaluation of the root cause of the problem with problem solving methods (such as an Ishikawa diagram, a *5-Why analysis* or a clustering approach), (II) gaining a first overview of the availability and the type of data and (III) a problem-oriented workshop with the users of the system to develop first insights into the nature of the problem. Working out topic (b) can be done with (I) an elaborated questionnaire provided to the potential team members, (II) conducting a structured interview, (III) an assessment done by the superior or (IV) a self-assessment of the team members.



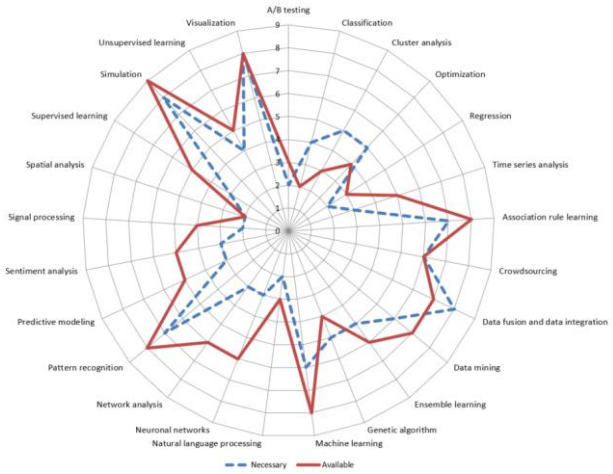


Figure 4 - Assessment of competencies

The polar coordinate diagram is the result of the assessment. The projects requirements and the competencies of each employee are visualized and an active decision whether an employee shall be allocated to a project can be made.

Based on the single diagrams the team can be set up; finding the optimal set up is an optimization problem.

Such problems can be solved with an add-in in Microsoft Excel, the Excel Solver. The Solver is a program that enables the user to find a certain value (maximum or minimum) for a target that is subjected to certain restrictions. The objective of the optimization is to set up a team with the best abilities to solve the big data problem.

Discussion

The higher motive of a company when setting up a big data and analytics teams is to create value for the company. This value does not have to be necessarily a monetary one in the beginning. There are other motives influencing this decision and finally leading to the founding of a big data and analytics team.

It is important to understand the effect that digitalization may have within the company in general and big data and analytics in particular in the value stream. Without knowing where the value stream will be affected, or even worse, not knowing where to apply big data and analytics within the value stream can result in a low return on investment or even in a loss for the company.

A common statement can be heard at conferences and symposiums quite often. It is that a company should “just” start with analytics, develop a “start-up mentality” and should not follow one path for too long and “fail fast” when the desired results cannot be achieved.

My opinion is a different one due to the fact that many companies just cannot develop a start-up mentality just hoping for a return on investment.

The success of big data projects origins partly in the correct planning, the correct application and the right execution of the mathematics. More important for the success of a big data project is the adequate organizational design, the definition and application of the right roles and competencies, and, at the end, an excellent management of the team and the project itself.

The results of research regarding leadership, motivation and social interactions apply also for a big data and analytics team. The Team needs to be managed accordingly, e.g. the cultural background, the generations perspective and their working style need to be taken into consideration by the management. The manager needs to able to lead such a team, and for this he has to have an understanding of the tasks, processes and methods the team is working with.

It can be argued that in most companies', there are already departments in which an analytics team could be integrated. Such departments could be IT, simulations, Research and Development or a continuous improvement department (CIP) that is working with Six Sigma methodology.

But those departments have a different focus; they provide structures, hardware or access to software in the best case. Those departments were not set up to be the driver for disruptive change in the analysis of mass data nor were they prepared to extract or treat or condition data following the requirements of a big data project.

Therefore, a different solution for the core team has to be found. Different scenarios were presented in this paper and allow the decision maker to work on the best solution for her or his company.

Supplement

The following supplements provides better insights into the scientific background to the literature research.

The extracts and conclusions are presented in a short version in the article.

Statistics and computer science technologies

The list shown here is a shortened version, taken from the original articles to provide an overview, and does not claim to be thorough. The list is in an alphabetic order (Manyika et al. 2011).

Technique	Explanation
A/B testing	A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing
Association rule learning	A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases (Agrawal et al 1993) (Hajek et al. 1966). These techniques consist of a variety of algorithms to generate and test possible rules
Classification	A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. Often described as supervised learning because of the existence of a training set
Cluster analysis	A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance
Crowdsourcing	A technique for collecting data submitted by a large group of people or community (i.e., the "crowd") through an open call, usually through networked media such as the Web (Howe 2006)
Data fusion and data integration	A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion
Data mining	A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression
Ensemble learning	Using multiple predictive models (each developed using statistics and/or machine learning) to obtain better predictive performance than could be obtained from any of the constituent models
Genetic algorithms	A technique used for optimization that is inspired by the process of natural evolution or "survival of the fittest." In this technique, potential solutions are encoded as

	"chromosomes" that can combine and mutate. These individual chromosomes are selected for survival within a modeled "environment" that determines the fitness or performance of each individual in the population. Often described as a type of "evolutionary algorithm," these algorithms are well-suited for solving nonlinear problems
Machine learning	A subspecialty of computer science concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data
Natural language processing (NLP)	A set of techniques from a subspecialty of computer science and linguistics that uses computer algorithms to analyze human (natural) language. Many NLP techniques are types of machine learning
Neural networks	Computational models, inspired by the structure and workings of biological neuronal networks (i.e., the cells and connections within a brain), that finds patterns in data. Neuronal networks are well-suited for finding nonlinear patterns. They can be used for pattern recognition and optimization. Some neuronal network applications involve supervised learning and others involve unsupervised learning
Network analysis	A set of techniques used to characterize relationships among discrete nodes in a graph or a network. In social network analysis, connections between individuals in a community or organization are analyzed, e.g., how information travels, or who has the most influence over whom
Optimization	A portfolio of numerical techniques used to redesign complex systems and processes to improve their performance according to one or more objective measures (e.g., cost, speed, or reliability)
Pattern recognition	A set of machine learning techniques that assign some sort of output value (or *label*) to a given input value (or *instance*) according to a specific algorithm
Predictive modeling	A set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome
Regression	A set of statistical techniques to determine how the value of the dependent variable changes when one or more independent variables is modified
Sentiment analysis	Application of natural language processing and other analytic techniques to identify and extract subjective information from source text material. Key aspects of these analyses include identifying the feature, aspect, or product about which a sentiment is being expressed, and determining the type, "polarity" (i.e., positive, negative, or neutral) and the degree and strength of the sentiment
Signal processing	A set of techniques from electrical engineering and applied mathematics originally developed to analyze discrete and continuous signals, i.e., representations of analog physical quantities (even if represented digitally) such as radio signals, sounds, and images. This category includes techniques from signal detection theory, which quantifies the ability to discern between signal and noise
Spatial analysis	A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data-set. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, e.g., addresses or latitude/longitude coordinates
Supervised learning	The set of machine learning techniques that infer a function or relationship from a set of training data (Cortes & Vapnik 1995)

Simulation	Modeling the behavior of complex systems, often used for forecasting, predicting and scenario planning. Monte Carlo simulations, for example, are a class of algorithms that rely on repeated random sampling, i.e., running thousands of simulations, each based on different assumptions. The result is a histogram that gives a probability distribution of outcomes
Time series analysis	Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data
Unsupervised learning	A set of machine learning techniques that finds hidden structure in unlabeled data
Visualization	Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses

Data analytics Algorithms

C4.5: The C4.5 algorithm is a classifier applied to generate decision trees, using the concept of information entropy. It was developed by Ross Quinlan, a computer science researcher. It is one of the most used tools in data mining.

k-means: The k-Means is an iterative method to partition a given dataset into a user-specified number of clusters, k. It is a type of unsupervised learning, which is applied to unlabeled data, e.g. data without defined categories or groups. Data points are clustered based on feature similarities (Trevino 2016).

Support Vector Machines (SVM): SVMs are supervised learning models that analyze data and recognize patterns; they can learn already from a short number of examples and develop at a fast pace. That algorithm was very popular in the 1990s.

Apriori: This algorithm is often applied to discover relations between variables in large datasets, using candidate generation.

Expectation Maximization (EM): The EM algorithm is suited to problems such as the estimation of the parameters of a probability distribution function (e. g. the estimation of the mean of a signal in noise). It produces maximum-likelihood (ML) estimates of parameters when there is a many-to-one mapping from an underlying distribution to the distribution governing the observation (Moon 1996).

PageRank: The PageRank algorithm was developed from Larry Page and Sergei Brin and provides the basis for the success of GOOGLE. It is a method to assign importance ranks to nodes in a linked database, such as any database of documents containing citations, the World Wide Web or any other hypermedia database (Page 1998).

AdaBoost: The algorithm deals with methods, which employ multiple learners to solve a problem (Dietterich 1997). The generalization ability with other algorithms to improve their performance is usually significantly better than that of a single learner. The contribution of the other learning algorithms is called the weak learner and is combined with the weighted sum to calculate the final output of the boosted classifier.

k-nearest neighbors (kNN): The k-nearest neighbor classification finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood.

Naïve Bayes: The naïve Bayes algorithm enables the researcher to construct a rule to assign future objects to a class, given only the vectors of variables describing the future objects.

Classification and Regression Trees: Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree (Loh 2011).

Roles within a Big Data Team

The list shown here is based on scientific articles (Debortoli et al. 2014, Pedersen 2017, Hitchcock 2017) and research in job portals (such as monster and stepstone).

Field of Application	Role
Analytics (proposed role: Data analytics expert)	Advanced analytics specialist Data analyst Analytics expert Statistician Mathematician Big data analyst Data scientist Web portal programmer
Data extraction, preparation and provision (proposed role: Data scientist)	Database administrator Data manager Software engineer Programmer Data engineer Analytics developer Application developer
Information Technology (IT) and hardware architecture (proposed role: Systems architect)	Platform specialist BI programmer Big data hardware architect Big Data IT system architect Information architect System architect
Software architecture (proposed role: Data software expert)	BI architect for Microsoft/Power BI, SAP/Business Objects, IBM/Cognos, QlikView, Cubeware, MicroStrategy, Teradata, Oracle, pentaho, and so forth Hadoop developer Hadoop applicant Data warehouse appliance specialist
Project and change management (proposed role: Data project manager)	Big data project manager Big data business consultant Big data business analyst Big data change agent Big data digital marketing

Bibliography

[1] Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. SIGMOD Conference, p. 207–16

[2] Adamson, D (1995). Blaise Pascal – Mathematician, Physicist and Thinker about God. ISBN 978-0-230-37702-8

[3] Baker, T. (2017). Performance Management for Agile Organizations. Palgrave Macmillan. Brisbane, Queensland, Australia. DOI 10.1007/978-3-319-40153-9

[4] Burton, R. M., Børge, O. 2018. The science of organizational design: fit between structure and coordination. Journal of Organization Design. Springer Open

[5] Chen, H., Chiang, R. H. L., Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly; Special Issue: Business Intelligence Research

[6] Constine, J. (2017). Facebook now has 2 billion monthly users…and responsibilities. https://techcrunch.com/2017/06/27/facebook-2-billion-users/

[7] Contractor, N. S., Wasserman, S., Faust, K. (2006). Testing multitheoretical, multilevel hypotheses about organizational networks: An analytics framework and empirical example. The Academy of Management Review, Vol. 31, No. 3, pp. 681-703. Stable URL: http://www.jstor.org/stable/20159236

[8] Cordova, A., Keller, K. M., Menthe, L.; Rhodes, C. (2013). Virtual Collaboration for a Distributed Enterprise. Chapter title: Conclusions and Recommendations. Published by RAND Corporation. Stable URL: http://www.jstor.org/stable/10.7249/j.ctt5hhw1p.13
[9] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning* 20(3). www.springerlink.com/content/k238jx04hm87j80g/
[10] Crosland, M. P. (1967). The Society of Arcueil: A View of French Science at the Time of Napoleon I. Harvard University Press.
[11] Debortoli, S, Müller, O., Brocke, J. von (2014). Vergleich von Kompetenzanforderungen an Business-Intelligence- und Big-Data-Spezialisten. Eine Text-Mining-Studie auf Basis von Stellenausschreibungen. Springer Fachmedien Wiesbaden. DOI 10.1007/s11576-014-0432-4
[12] Dietterich T. G., (1997). Machine learning: Four current directions. Department of Computer Science, Oregon State University, Corvallis
[13] Dijcks, J.-P. (2013). Big Data for the Enterprise. An Oracle White Paper. Oracle Corporation. http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf
[14] Donald A. M., Peppard, J. (2013). Why IT fumbles analytics. Harvard Business Review. http://hbr.org/2013/01/why-it-fumbles-analytics
[15] Economist, The (2011). Beyond the PC. Special report on personal technology. http://www.economist.com/node/21531109
[16] Ebert, P., Freibichler, W. 2017. Nudge management: applying behavioural science to increase knowledge worker productivity. Journal of Organization design. Springer Open
[17] Geissbauer, R., Schrauf, S., Berttram, P., Cheraghi, F. (2017). Digital Factories 2020. Shaping the future of manufacturing. Published by PricewaterhouseCoopers GmbH Wirtschaftsprüfungsgesellschaft (PwC)
[18] Hackl, P. & Katzenbeisser, A. (1996). Statistik. Oldenbourg, München, Wien; 10. Edition
[19] Hadjinicolaou, N., Dumrak, J., Mostafa, S. (2018). Improving Project Success with Project Portfolio Management Practices. Springer International Publishing AG
[20] Hajek, P., Havel, I. and Chytil, M. (1966) The GUHA method of automatic hypotheses determination. *Computing* 1(4), p. 293–308
[21] Håkonsson, T., Carroll, T (2016). Is there a dark side of Big Data – point, counterpoint. Journal of Organization Design. Springer Open
[22] Hitchcock, E. (2017). 5 Big Data Job Descriptions to Hire an All-Star-Team. https://www.datameer.com/company/datameer-blog/big-data-job-descriptions-hire-recruit-team/
[23] Howe, J. (2006). The Rise of Crowdsourcing. *Wired*, Issue 14.06, June 2006
[24] Hudec, M. & Neumann, C. (n. d.) Was ist Statitik? Geschichte, Grundlagen, Anwendungen. Institut für Statistik der Universität Wien. http://www.stat4u.at/download/1417/WasIstStatistik.pdf
[25] Kübler, R. V., Wieringa, J. E. & Pauweis, K. H. (2017). Advanced Methods for Modeling Markets. International Series in Quantitative Marketing. Chapter 19: Machine Learning and Big Data, p. 631. Springer International Publishing AG.
[26] Kuls, N. (2018). Absturz einer Internet-Ikone. http://www.faz.net/aktuell/finanzen/facebook-datenskandal-absturz-einer-internet-ikone-15515647.html
[27] Larson, E. (1989). How do they get your name? Direct-mail firms have vast intelligence network tracking consumers. http://articles.orlandosentinel.com/1989-07-26/lifestyle/8907254531_1_subscribe-to-magazines-subscriber-list-junk-mail
[28] Loh, W.-Y., (2011). Classification and regression trees. WIREs Data Mining and Knowledge Discovery. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8
[29] Lohr, S. (2013). The Origins of 'Big Data': An Etymological Detective Story. https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/
[30] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
[31] Marr, B. (2016). A short history of machine learning. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#4eb8a56315e7

[32] Mauro, A. D., Greco, M., Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. http://dx.doi.org/10.1063/1.4907823
[33] Milutinovic, V, et al. (2017). DataFlow Supercomputing Essentials. Algorithms, Applications and Implementations. Chapter 5: DataFlow Systems: From their origins to future Applications in Data Analytics, Deep Learning, and the Internet of Things. Springer International Publishing AG
[34] Monica, P. R. L. (2018). Facebook has lost $80 billion in market value since its data scandal. http://money.cnn.com/2018/03/27/news/companies/facebook-stock-zuckerberg/index.html
[35] Monnappa, A. (2017). How Facebook is using Big Data – The good, the bad, and the ugly. https://www.simplilearn.com/how-facebook-is-using-big-data-article
[36] Moon, T. K. (1996). The expectation-maximization algorithm. Elect. & Comput. Engineering Department, Utah State University, Logan, USA
[37] Moore, D. (1992). Statistics for the twenty-first century. Teaching statistics as a respectable subject. The Mathematical Association of America, Washington, DC
[38] Nunan, D. & Domenico M. D. (2015). Big Data: A normal accident waiting to happen? Journal of Business Ethics. Springer Science+Business Media Dordrecht 2015
[39] Page, L. (1998). Method for node ranking in a linked database. Patent number: US19980004827 19980109
[40] Panson (2015). Top Data Mining Algorithms Identified by IEEE & Related Python Resources. https://www.datasciencecentral.com/profiles/blogs/python-resources-for-top-data-mining-algorithms
[41] Pedersen, C. L., Ritter, T. (2017). The 4 types of project manager. Article project management. Harvard Business School Publishing Corporation. Reprint H03SJ5.
[42] Quinlan, J. R. (1986 and 2007). Induction of Decision Trees. Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney, Australia
[43] Reimer, K., Rutz, O. J., Pauwels, K. H. (2014). How online consumer segments differ in long-term marketing effectiveness.
[44] Rosenthal, C. (2013). Big data in the age of the telegraph. https://www.mckinsey.com/business-functions/organization/our-insights/big-data-in-the-age-of-the-telegraph
[45] Russom, P. (2011). Big Data Analytics. TDWI Best Practices Report, fourth quarter 2011. https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf
[46] Scott, J. (2014). 5 Google projects that changed big data forever. https://mapr.com/blog/5-google-projects-changed-big-data-forever/
[47] Soni, Y. (2017). So what is machine learning? https://becominghuman.ai/machine-learning-for-dummies-explained-in-2-mins-e83fbc55ac6d
[48] Statista.com (2018). Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions) (2018). https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/
[49] Sullivan, D. (2015). How machine learning works, as explained by Google. https://martechtoday.com/how-machine-learning-works-150366
[50] Teece, David J., Linden, G. (2017). Business models, value capture, and the digital enterprise. Journal of Organization Design. Springer Open
[51] Tilly, C. (1984). The old new social history and the new old social history. Research Foundation of State University of New York for and on behalf of the Fernand Braudel Center. Stable URL: http://www.jstor.org/stable/40241514
[52] Trevino, A. (2016). Introduction to k-means clustering. https://www.datascience.com/blog/k-means-clustering
[53] Vagata, P. & Wilfong, K. (2014). Scaling the Facebook data warehouse to 300PB. https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/
[54] Wasserman, L. (2012). Normal Deviate. Thoughts on Statistics and Machine Learning. Statistics versus machine learning. https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/
[55] Wehrt, K. (1985). Beschreibende Statistik: Eine Einführung. Campus-Verlag, Frankfurt
[56] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., Steinberg, D. (2007). Top 10 algorithms in data mining. Springer Verlag London Limited