Big Data Management

Innovation potential analysis for the new technologies for managing and analyzing large amounts of data

Executive Summary

Technische Universität Berlin

Technische Universität München

Westfälische Wilhelms-Universität Münster

Prof. Volker Markl Prof. Helmut Krcmar Prof. Thomas Hoeren

Commissioned by

Federal Ministry of Economics and Energy

Abstract

The increasing degree of networking in the Internet of Things and Services and the use of advanced sensor technology and simulation models in Industry 4.0, in service companies, in research and in the private sector are resulting in ever greater data availability. The analysis of this data will revolutionise commercial, scientific and social processes by providing rapid and comprehensive data-driven support for decision-making. Companies in particular can expect to gain significant competitive advantages. This trend is currently described by the terms "big data" or "data science". Here, big data signifies that both data and analyses based on this data have attained a new quality and complexity in recent years.

If we are to ensure the lasting competitiveness of Germany as a business location, and to continue to generate innovations in the modern information society, we will need a coherent interplay of four fields:

technology: provision of effective methods and tools to analyse large quantities of heterogeneous data at a high data rate,

commercial exploitation: creation of applications to develop new markets or to strengthen existing markets,

legal framework: ownership of data, data protection law, copyright and contractual and liability problems,

and the **training of skilled workers.**

The development of ICT for big data management is a logical priority of the Federal Ministry of Economics and Technology, and one which particularly merits support.

This executive summary is an abridged version of a detailed study which evaluates this interplay in order to move, via management and analysis, from big data to smart data. The complete text of the study (in German) is available at `http://www.dima.tu-berlin.de/menue/research/big_data_management_report/`.

1

Introduction: What are the New Big Data Qualities?

Big data is characterised by a new level of complexity in terms of the data and the analysis conducted on the data. This new type of data complexity is characterised by requirements in terms of volume, velocity, variety, and veracity, which cannot be met by conventional database systems. For example, the analysis of big data requires:

- the storage and processing of enormous quantities of data,
- whilst the window for decision-making within which the results of the analysis have to be provided becomes smaller and smaller,
- a large number of different data sources (e.g., time series, tables, text, images, audio and video streams) to be included in the data analysis and
- due to the fuzziness of some data sources (e.g., sensors with a fixed precision level) or of information extraction procedures and integration procedures, systems and analysts need to handle probability-based models and confidences.

Also, new declarative languages are needed for specifications and the automatic optimisation and parallelisation of complex data analysis programmes (including new statistical and mathematical algorithms) in order to cope with the volume, the processing speed, the different data formats and the reliability of the data. Furthermore, big data involves a new complexity of analysis, as reflected in the fact that models are generated from the data in order to support decision-making. This necessitates the use of advanced data analysis algorithms drawn from statistics, machine learning, linear algebra and optimisation, signal processing, data mining, text mining, graph mining, video mining, and visual analysis.

These requirements will result in a paradigm shift in data analysis languages, data analysis systems, and data analysis algorithms, making entirely new types of applications possible.

2 Big Data Management: an Opportunity for Innovation in Europe

The new challenges brought about by big data represent a great opportunity for German and European companies, both in terms of technology and in terms of applications for this field. Existing products in the commercial database market, which is largely dominated by U.S. firms, are based on technologies, which cannot cope with big data because of a lack of scalability, lack of error tolerance, or restricted programming models. This means that there is a new international situation in the field of scalable data analysis systems. Germany is well positioned here. After the U.S., Germany has the second-strongest research community in the field of scalable data management. This community is already engaged in a large number of activities in basic research centered on big data (e.g., the Stratosphere System at Technische Universität Berlin, Hyper at Technische Universität München, the research on Hadoop++ and HAIL at Saarbrücken University). In addition to an open source movement, which is also active in Germany, many companies, and particularly start-ups are challenging established providers like IBM, Oracle and Microsoft.

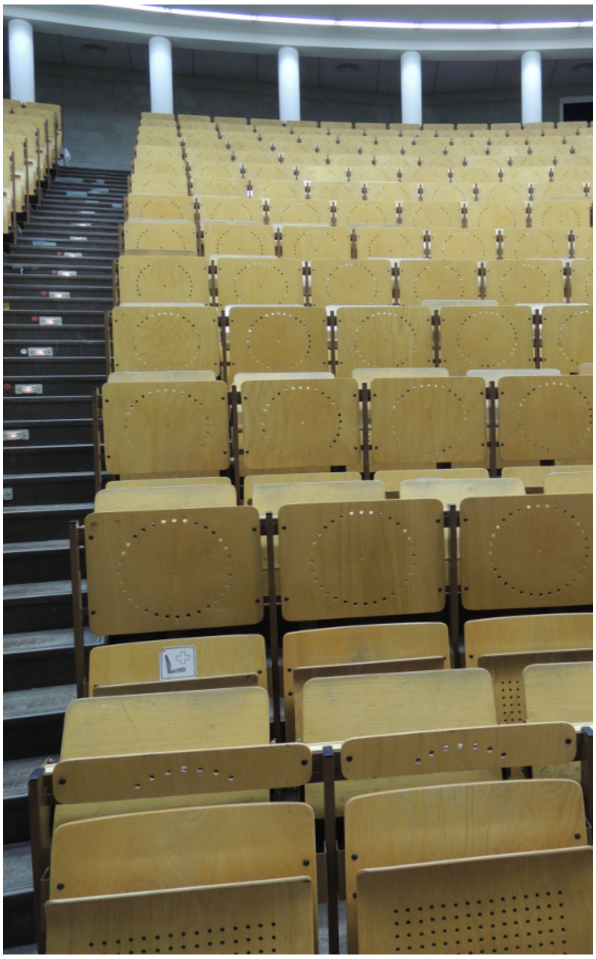
Among the challengers pursuing opportunities and seeking market share in the emergent big data market there are a host of German technology companies as well as large firms (e.g., SAP and their HANA product, Software AG and their Terracotta product) and many high-tech start-ups (e.g., ParStream and Exasol). If the product and marketing strategies of these companies are to be supported, it is important to create a climate of technology transfer and innovation, which enables German firms and particularly SMEs, university spin-offs and startups in the field of scalable data processing and data analysis to compete as equals with the startups particularly found in Silicon Valley, in the UK, and in China.

In this way, policy-makers can make an important contribution towards the competitiveness of German firms in the future development and commercialisation of key enabling technologies for big data, and pave the way for Germany to enter the billion-dollar big data market via scientific achievements and innovations.

3 Training at Higher Education Institutions

In order to optimally prepare German commerce, science and society for this global trend, there is a need for highly coordinated research, teaching and technology transfer activities in the field of data analysis and scalable database management. Big data is no longer just a challenge for a specific sector; rather, it affects all areas of the economy, all organisations and all users of digital technologies. The novel job description of the "data scientist" combines knowledge of data analysis procedures (e.g., statistics and machine learning, optimisation, linear algebra, signal processing, language processing, data mining, text mining, video mining, and image processing) with technical skills in the field of scalable data management (e.g., database systems, datawarehousing, information integration, distributed systems, computer networks, and computer architectures) and practical systems implementation skills.

This sort of training should be backed by practical application-based projects to teach skills in certain application areas.



4

Legal Aspects of Big Data

The technical developments in recent years have substantially increased the available quantities of data. These newly introduced technologies and their applications not only enable data storage for a virtually unlimited amount of time, but also to analyse the data in greater detail, for example, in terms of user behavior patterns. However, the volume of data also raises numerous legal questions, mostly related to the fields of data protection-, copyright- and contract law. Furthermore, there is an ongoing legal debate, which when resolved will have serious consequences for the whole big data industry. The question whether there is a right of **data ownership right**, and if so who that would be.

Data Ownership

A property Data rights for data might seem a dispensable theoretical construct unnecessary given the already existing ownership, copyright (e.g., attributed to databases) and privacy protection. But data itself has become an economic factor in its own right. It represents merchandise which has an independent tangible value. With this backdrop it seems appropriate to develop an absolute protection regime.

Dogmatically, the concept of data ownership creates challenges. Proprietary data assumes that data can be clearly attributed to a legal subject and thus grants the subject full property rights in return. A classification according to Section 90 ff. of the Civil Code (BGB) seems impossible. The datum itself is not a tangible asset according to Section 90 of the Civil Code (BGB) rather it is physically dependent. To include a datum as major component of the data medium according to Section 93 of the Civil Code (BGB) would conflict with the view that in this case "ownership" and property rights are inseparable. But this needs to upheld to appropriately compensate for damages, for example, attributed to a data base.

This is exemplified in referring to a concrete situation in which the economic damage did not occur to the owner of the data medium, but rather to a third party using the data for economic benefits. While it was possible 25 years ago to recourse to property rights when solving a legal case involving data loss on a personal data medium, this construct is no longer applicable due to modern technological advances in storing data in the "cloud". The formerly argued "right to personal data stock" in these circumstances is "another right" according to Section 823 Civil Code (BGB) that is potentially viable, but is met with dogmatically objections. To qualify as an exclusive right, it would need to offer an exclusion function similar to the rights defined in section 823 Abs 1 Civil Code (BGB). It is unclear who this exclusion right should be attributed to, thus this solution could only serve as an auxiliary construct that creates more problems than it solves.

Assigning data ownership rights to legal subjects could also be achieved by means of data protection rights. This only creates a legal responsibility for data and shall not be understood in such a way as data subjects holding exclusive rights on individual data records in the legal sense of a property. The same holds true for a potential solution of using the exclusive sui-generis right on databases. This demonstrates an investment protection that protects from economic exploitation by others, but does not extend to prevent linking data and people. The linkage problem however can be solved by appealing to criminal law, as Section 303a Criminal Code (StGB) explicitly protects data. Linking the subject of protection to a legal entity is achieved through the scripture act (Skripturakt), thus by means of the technical creation process itself. Transferring this concept to civil law enables a definite attribution and thus establishes the possibility for "data ownership".

It can be concluded that the derivation of an ownership on data is dogmatically possible. But it is unforeseeable whether the institution of data ownership will prevail as the discussions about it are just starting. Nevertheless attention should be devoted to this topic especially from the point of view of big data companies. Depending on the outcome of the debate the "data owner" assertions could affect the data record handling, independently of any data protection or copyright claims. Furthermore, a clarification to the question of data ownership would also have dramatic impacts on bankruptcy law related issues as well.

Data Protection

Currently ongoing scientific discussions vehemently point out the conflicts that exists between big data and data protection. Looking at the data protection principles such as appropriation, transparency, direct survey, data reduction and data economy, and the principle of prohibition with the reservation of permission it becomes clear that big data can collide with the convention for the protection of individuals with regards to automatic processing of personal data. To prevent data protection breaches big data companies must therefore consider a manifold of problem areas during decision making.

One problem is already the question of applicability of the German privacy law. The generally valid principle of territoriality already causes difficulties, if data is distributed worldwide onto different sites, the data is ephemeral and can also very rapidly change its determined location. Big data companies must therefore consider a multitude of different legal systems when dealing with data.

Furthermore the permissibility of dealing with data from a copyright point of view depends on an approval by the data subject or the determination of legitimacy by statute attributed to the processing agency. This principle of prohibition with the reservation of permission is increasingly criticised as being unsuitable since with the ubiquity of data processing in, for example, smartphones almost everyone could potentially become a data processor. The protection of informational self-determination is placed above all at the beginning, to subsequently define many in parts very extensive statutory permissions. It is suggested to rethink the concept of prohibition with the reservation of permission in the scope of modernizing the data protection law. In this way big data companies could be relieved from the complex task to determine the "appropriate" legitimate elements of a rule.

In this context the general question of the effectiveness of the institution of "consent as legitimate factor" arises. On the one hand there are existing concerns towards the optional nature of the consent, if it becomes the de facto good in return for a "free" service and thus develops into merchandise itself. On the other hand it needs to be checked how consent can be obtained legally and whether its effective duration could be limited.

Altogether the representative scenarios presented and a multitude of other problems show that data protection laws are an obstacle for big data companies. On the one hand this is the desired effect to successfully protect personal data. On the other hand it can be seen that the principles of data protection originated from an era that didn't cater to the big data phenomena. For that reason modernisation proposals are needed which balance the opposing interests in an adequate manner.

Copyright

Dealing with big data creates a lot of copyright related questions as well. A universally valid assessment of copyright related questions which holds true for any big data solution is impossible to achieve as concrete problems always arise as consequences of the individual implementation of the processes. Protection of copyright can exert effects along two different dimensions. On the one hand it is possible that the data processing company itself could claim copyright, in particular the sui-generis database protection. On the other hand dealing with data itself infringes upon copyrights or related protected privileges of third parties. Especially the latter needs to be considered by big data companies, when making concrete data processing decisions.

This context, similar to the data protection discussion, yields the question of applicability of the German statue. The principle of territoriality in accordance with the German law is generally applicable to big data solutions that are used in Germany. Internet related cases which cannot be attributed to a concrete territory lead to conflict of laws questions that have yet to be answered completely. Big data companies in turn are faced with an uncertain legal environment since multiple different legal statues could be applicable.

Copyright protection can only be guaranteed if a sufficient level of originality is present in the work. Individual data records regularly do not present the necessary individual imprint, and as such copyright infringement are not to be expected. But something contrary could hold if user generated content, for example derived from a social network is to be analyzed. Such records could be individually protected as photographic images (Section 72 UrhG) photographic works (Section 2 Abs. 1 Nr. 5 UrhG), or literary works (Section 2 Abs. 1 Nr. 1 UrhG). An analysis would therefore touch upon the right of reproduction and the right to make available to the public.

The permissibility of data handling from the point of view of copyright law dependends on whether existing statutory exceptions apply in favor of the data processing company or the rights holder is acknowledging the handling. Statutory exceptions will generally not authorize the handling of large amounts of data. This finding serves as a basis for discussion as to whether the copyright act (UrhG) should be expanded with further exceptions to accommodate the new dimensions of data flows. Without "suitable" exceptions big data companies can only resort to seeking approval from the respective rights holders. This however bears considerable practical problems. New technologies in the digital age lead to the fact that almost inevitably many foreign exclusive rights are touched upon so that a great number of contractual agreements would be required. To avoid this problem the figure of factual consent is introduced. Whether such a construct would in fact be viable for authorizing the processing is still undecided. It can be concluded that there is a clearly visible trend that in the case of exercised copyrights concerning big data is only possible under unfavorable conditions.

This statement could be further emphasized by the fact that another copyright protection act could impede big data processing as well. If the data sets are sources from a third party database, the sui-generis database protection according to Section 87a ff. UrhG could conflict with the actions of the big data company. Sui-generis protection is independent of the level of originality. It rather protects a database if a substantial investment has gone into the creation. When exactly this condition is met is determined on a case by case basis. This results in turn to uncertainties that need to be addressed before managerial decisions are undertaken. If the condition is met this implies that without consent of the database owner only insignificant parts of the dataset can be used for the data processing. Databases that fall below the regulated threshold in turn can be used without exceptions.

Overall these sketched problems prove that copyright with all its unanswered legal issues as well as the regulations that are partly in need of modernization could present a barrier for the big data sector. In any event the focus on data privacy law from the point of view of a business, should not blur the view on corresponding copyright related regulations.

Contractual and Liability Issues

Further uncertainties for big data companies arise in regards to the contract terms as well as liability issues. Contract design is complicated due to the many technical aspects involved as well as the existing touch points with data protection and copyright law. It is hard to make generally applicable contract law statements. Nevertheless it can be asserted that the regulations in terms and conditions, which exclude or restrict liability for data loss or data damage are in danger of being ceased during a global juridical review of the terms and conditions. Furthermore there are general issues with liability of big data companies. This presupposes misconduct on part of the respective company. This could be owned to the circumstances of transmitting faulty information. It is still not sufficiently defined which criteria need to be present to declare such a case.

Conclusion

The development of big data in Germany is significantly influenced by the established law. Of great importance will be the ongoing debate about a potential "data ownership". In addition there are some existing barriers that impede especially data privacy as well as copyright regulations. Part of a debate on a modernisation of these fields of law should include whether these regulations are still up to date and where appropriate alterations are possible which would appeal to economic interests of data processing companies without neglecting the legitimate interests deserving protection of data subjects. As long as this process is ongoing, big data companies must deal on a case by case basis with the fact as to whether their big data solution operates within the legal bounds as defined by the data privacy as well as data copyright law.

5 The Innovative Potential of Big Data

Our study investigated the innovative potential of big data both in select vertical market sectors and on across horizontal markets. This chapter summarises our findings.

We only relied on big data studies which contain information about Germany, have quantitative data and contain comparable statements, so that they can be linked up together. A total of nine studies were evaluated. The sample size of the studies considered here (i.e., BARC, BITKOM, Computing Research (two studies), Experton Group, Fraunhofer IAIS, Interxion, TATA Consultancy Services, TNSinfratest) was large at n > 4492.

The core messages of the studies can be summarised as follows:

1. Big data has thus far chiefly been a topic for IT experts, but has been relatively unknown beyond that.

2. Big data will alter company management, business processes and information logistics.

3. The main need for specific action at present is perceived in the field of information logistics.

4. Companies are only just starting to develop strategies for the use of big data reaching beyond the analysis of structured data.

5. There is a lack of skilled workers, organisational structures, and processes in companies to use big data. These will be in even greater demand in future.

6. As companies take a greater interest, the big data market will grow considerably.

Results of the Empirical Study. In addition to the above-mentioned analysis offered existing studies on the innovative potential of big data, an independent empirical study (with 185 participants) was carried out. The aim of this study was to validate and supplement the existing findings.

The investigation was carried out in the 2nd and 3rd Quarter of 2013. The total amount of participants was 185.

Background of the Participants. Figure 5.1 shows the professional background of the participants of the study, 16% of the participants were decision makers. The majority of the participants had a background in computer science or related fields (Figure 5.2).

Most participants were between 20 and 54 years of age. (representing 164 responses and 16 participants were older than 54) and attributed themselves mostly to the information technology sector.

Out of the 90 participants who identified as part of the private sector, 40 were

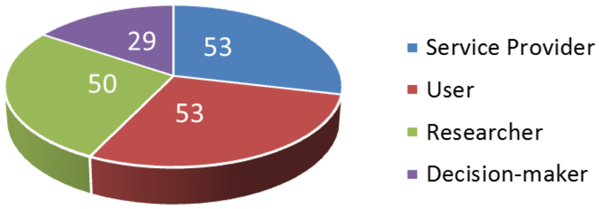


Figure 5.1: The field of work of the participants in the study

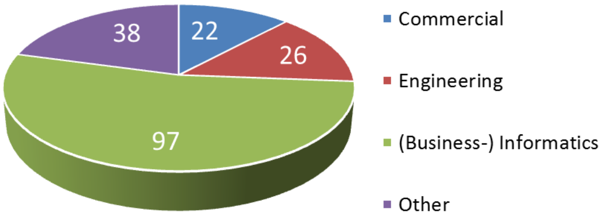


Figure 5.2: The education of the participants in the study (Answers=183)

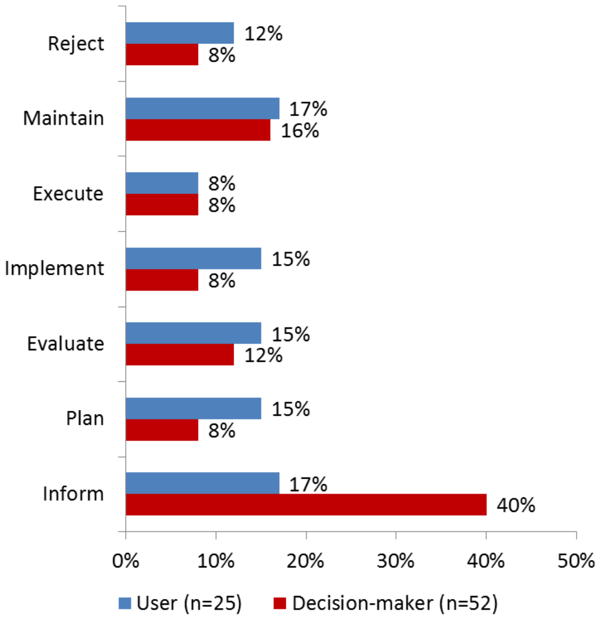


Figure 5.3: State of Big Data Projects

part of small and medium sized enterprises (SMEs). Exactly 47 participants came from organizations with more than 1000 employees, 23 from organizations with 2500-1000 employees and 39 from 250 or less employees (109 responses in total). Precisely, 38 companies had an annual turnover of more than 250 million euros, 66 had a lower annual turnover, out of which 24 had an annual turnover between 5 and 50 million euros. The next sections discuss the results of the study in detail.

Big Data Projects are Still in a Very Early Stage. About 40 % of the decision-makers state that they are still in an information phase regarding big data technologies (see Figure 5.3) Only 8 % of decision-makers have already considered implementing those technologies. A quarter of decision-makers stated, that they are already planning and/or testing Strategies, Roadmaps and Measures, including cost-benefit analyses. On the side of practitioners, things have already progressed further. 15 % have already implemented a big data strategy. Furthermore, more companies are in the phase of implementation, testing and planning.

Big Data Will Be Important in 5 Years. Decision-makers and practitioners have been asked to assess the importance of big data in practice as part of this study. While 44 % of the practitioners (n=50) rate the topic of big data as very important or rather important, only 29 % (n=24) of decision makers did so. For the year 2014, 61 % of practitioners and 43 % of decision makers had this view. However the biggest importance of the topic of big data was asserted for the next five years (78 % of practitioners and 60 % of decision makers).

Big Data Will Have Established Itself in Less Than 10 Years. A large majority of the participants expects that big data will prevail everywhere in their industry in less than 10 years. This assessment is shared by providers and decision makers, as is shown in Figure 5.4.

Big Data Has High Value Creation Potential. About 50 % of decision-makers and 72 % of practitioners assess the value creation potential of big data as high or very high. Only 25 % of decision-makers and 8 % of practitioners assess the value

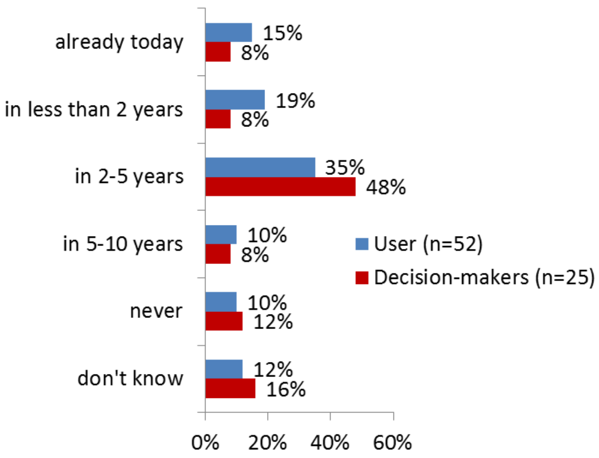


Figure 5.4: When is Big Data starting to provide a competitive advantage?

creation potential of big data as low. It is noteworthy that practitioners assess the potential of big data significantly more positive than decision-makers.

Measurable Value Attributed to Big Data? About 16 % of decision makers and 19 % of practitioners have stated that they already managed to create measurable value with big data. But 72 % of decision makers and 54 % of the practitioners stated that they could not yet estimate the actual value of the value contribution. These results indicate the discrepancy between the value creation potential and the actually measurable contribution.

Besides Relational Data, Transaction Data, Text Analyses and Web Analyses Dominate. The providers of big data technologies are highly diversified and can provide analysis tools for many different data types. However predominantly tools for relational data and highly structured data such as transaction data are requested. Video, image, and audio data are rated as rather unimportant (see Figure 5.5).

From a practitioners' point of view, analysis of transaction and web data are the major concern regarding already carried out analyses. It is particularly interesting, that only few practitioners plan the future application but rather do not yet know whether such analyses should actually be carried out at all.

High Quality Data , Lack of Skilled Personnel and Lack of Economic Feasibility are the Biggest Challenges. Decision makers identified the lack of incorporated data analyses into actual decisions as the largest obstacle to the adoption of big data. Decision makers also voiced major concerns regarding the legal framework of privacy. The shortage of skilled employees for data analyses is also named as a concern. Practitioners also largely see privacy issues as the largest concern. They share the view that incorporating data analyses into the decision making process is a major challenge. Providers on the other hand largely see privacy concerns as the major challenge. Providers share the challenge to demonstrate how data analyses can be incorporated into the decision making process. Furthermore providers struggle with the complexity of the structure of data in enterprises. In summary, the largest challenges of big data lie in a value adding integration of big data into decision making processes, the lack of skilled data analysts and concerns regarding privacy.

The results of the survey can be summarised as follows:

1. Big data is being viewed in the light of more effective corporate decisions. However, providers and users are facing the challenge of developing convincing combinations of data analysis, commercial decisions, and value. There is an opportunity here in in-house processes, since a large amount of data has not been used so far.

2. German providers of big data technology are capable of meeting the needs of German users. However, a key challenge is the elucidation and explana-

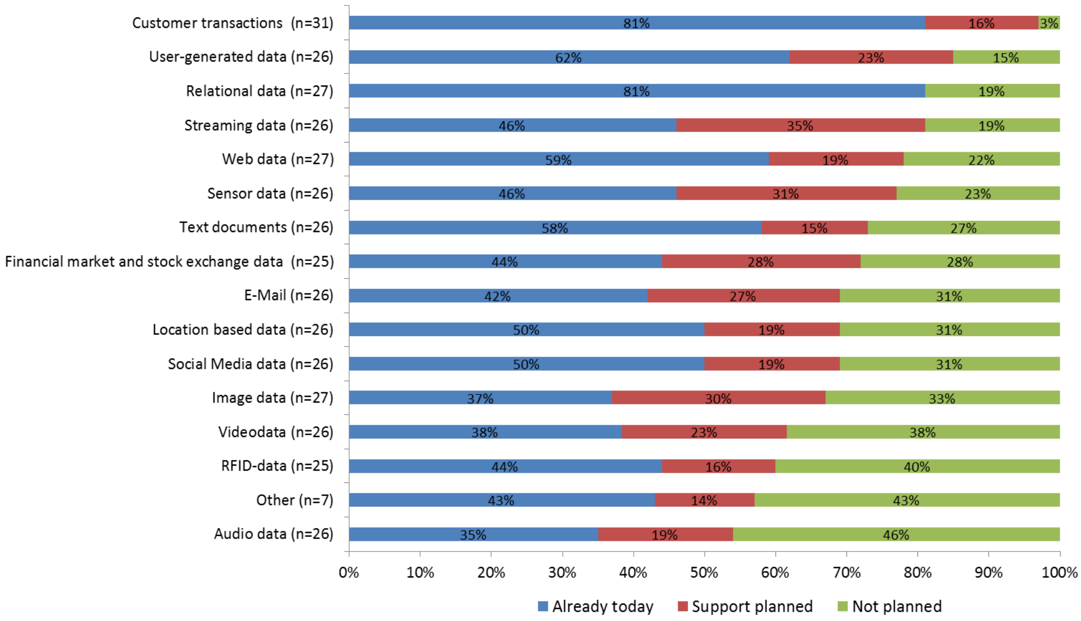


Figure 5.5: Service provider perspective: Existence of Big Data tools for supporting various data types

tion of data protection rules and the development of viable commercial solutions.

3. Also, big data is believed to offer potential in the field of new business models, products and services. User companies are still unclear about what data analyses are of relevance and value for their respective business processes. Here, providers and users could work intensively together in order to develop corresponding business models, products and services.

4. One important challenge lies in the responsible use of personal data. All of the stakeholders can see major challenges in this field. Here, it is necessary not only to bring the law into line with the state of the art, but in particular to educate and to manage expectations.

5. Another key challenge lies in the availability of appropriately trained staff. There is a need here for suitable training courses in order to provide companies with the necessary expertise.

6. Targeting the survey at small and medium-sized firms reveals that there are hardly any differences in the perception of the innovative potential. This can be interpreted as an opportunity for small and medium-sized firms, since there are clearly no additional hurdles for this business sector. Further to this, the studies considered and the authors' own study discuss general horizontal potential as sustainably classified individual sectors.

Cross Sector and Sector Specific Innovation Potential. For the analyses of strengths, weaknesses, opportunities and threats (SWOT), literature research was carried out in order to assess the individual factors. Different strengths, and weaknesses as well as opportunities and threats could be derived. The results are shown in Figure 5.6.

Based on the reviewed studies as well as the own study sectors with an outstanding innovation potential could be identified, these are:

SWOT-Analysis		Internal Analysis	
		Strengths	Weaknesses
External Analysis	Opportunities	• Fast and accurate analysis • Better strategic decisions • Improved control for operational processes • Better analysis with Business Intelligence • Cost reduction • More flexible services, targeted marketing campaigns • Better information for internal risk management • Better understanding of the market • Improved customer service	• Monitoring leads to an increase of internal control • A reduction of the data set led to unanalyzed data points • Big Data creates more trustworthy data • Customer specific solutions (transparent customer)
	Threats	• Distribution Data • Fraud Detection • Unstructured/informal communication • Lack of technical and domain knowledge	• Privacy • Data integrity and security • Social acceptance • Ethics • A lack of plausible and convincing application scenarios • Technical problems • Costs

Figure 5.6: SWOT Analysis for the Assessment of the Potential of Big Data

- public sector,
- Industry 4.0,
- health sector/life sciences,
- market research, (social) media and entertainment,
- mobility services,
- the energy sector, plus
- risk management and the insurance industry.

The studies and the selected sectors show that the data volume is manageable via various technologies, such as Hadoop, Stratosphere, in some cases also via SAP HANA, Par-Stream, or other in-memory databases. Data protection and data security offer considerable potential for big data technology in Germany. Due to the fact that Germany has a well-developed understanding of trusted data, Germany can become a market leader in this field. To this end, there is further need for research in the field of data protection and data security, particularly in terms of the integration of data protection functionality into existing or emerging data analysis systems or algorithms. The protection of data is of considerable significance, particularly in order to uphold commercial interests.

6

Analysis of Big Data and Big Data Technologies

On the basis of the requirements of industry, as sketched out in the preceding chapter, one can derive four core requirements for the management of big data.

1. Handling of large, heterogeneous quantities of data
2. Complex data analysis algorithms
3. Interactive, in many cases visually supported data analysis
4. Comprehensible data analysis.

Existing data management systems cannot cope with these challenges. In order to master them, there is a need to develop scalable, easy-to-use data analysis systems and new algorithms or paradigms for data analysis which address the various aspects and requirements at the same time.

Analysis of Big Data

A new level of data complexity reflects itself in novel requirements regarding volume, velocity and veracity, which are not covered by existing database solutions. The analysis of big data requires the storage and processing of massive data sets in the terra- or even petabyte range. At the same time the timeframe for decisions, in which analysis results have to be generated, becomes smaller and smaller. Data analysis systems have to provide analyses with low latency even when facing high data rates with which new data sources have to be integrated into the database. Furthermore, a lot of different data sources each containing data of different structure (such as time series, spreadsheets, text, images, audio and video streams) are tapped for analyses.

A new level of complexity regarding the analyses of big data manifests itself in the fact that models for decision support have to be generated from the data. This requires the application of advanced data analysis algorithms, in particular methods of statistical machine learning, linear algebra, optimization, signal processing, data mining, text mining, graph mining, video mining as well as visual analyses. Therefore the data analysis system has to process complex algorithms from linear algebra, statistics, or optimization theory in a timely manner. These algorithms are characterized by a combination of user defined functions (UDFs), iterative, status algorithms and the common operations of relational algebra. This combination of iterative algorithms and user defined functions is not provided by either traditional SQL-based database systems nor existing big data solutions (e.g., Hadoop, Pig, Hive, Storm, Lambda Architecture, etc.).

Thus the development and commercialisation of modern data analysis systems that combine relational data processing with algorithms of statistical machine learning provide a high innovation and market potential.

Technical Challenges of Big Data Technologies

The interactive and iterative data analysis process necessitates the mastering of three technical challenges:

1. People must be able to describe the desired result of the inquires in a high-level language,
2. the technology must process iterative data flows and
3. the technology must also be able to process unknown programmes of other providers, known as user-defined functions (UDFs), sufficiently quickly.

Programming Models for the Rank and File Analysts. Besides to Hadoop a multitude of interesting research work on massively parallel processing of data driven, iterative algorithms exist. However, none addresses the development of a declarative specification and automatic optimization of iterative algorithms. Therefore, the analysis of big data requires skills in distributed systems programming, knowledge in the domain of analysis as well as a solid understanding of machine learning methods. People with such a combination skills are quite rare. Overcoming this shortage will be the critical success factor not just for new big data technologies but also for a broad application and uptake of big data analytics.

Iterative Data Processing. Iterative data analysis methods compute the result of an analysis in a lot of individual steps. Each step generates an intermediary result or intermediary state. Since, such computations have to be carried out in parallel due to the high data volume, the repeatedly updated state has to be efficiently distributed, stored and managed across many machines. For efficiency reasons, the state has to be held in main memory. Many algorithms also require a lot of iterations until they converge to the final solution. It is thus imperative to compute iterations with low latency to minimize the total query time. In some cases the computational effort decreases significantly after the first few iterations. Systems such as MapReduce and Spark carry out all computations in each iteration, even if part of the result has already converged and no longer needs to be updated. Contrary to this approach, real iterative data flow systems like Stratosphere or specialized graph processing systems such as GraphLab or Google's Pregel can exploit this property and reduce the computational effort for each iteration.

Processing of User Defined Functions with Low Latency

User defined functions are not inherently part of the system, they are a long known and supported concept in relational data base systems. However interfaces are often too restrictive to allow the implementation of complex algorithms. Google's MapReduce, SCOPE, Stratosphere and Spark are the only systems that offer more expressive user defined functions that can be executed in parallel. The degree of parallelism is usually determined by the semantics of the programming interface (e.g., second order functions like map and reduce in Hadoop or further functions like match, cogroup or cross in Stratosphere).

User defined functions do not belong the functional scope of the system but are often executed externally. For that reason, their usage generally leads to higher execution costs. These costs can be lowered, when the UDFs are deeply embedded inside the system. Another challenge is the treatment of UDFs during the optimization of data analysis programs. The semantics of a UDF generally are not known to the system, data analysis programs with UDFs can only be optimized if the optimizer has additional information. First approaches attempt to obtain this information by manual annotations or static code analysis.

State of the Art

The new challenges posed by big data are not covered by established systems. Thus, there is a chance to break the quasi monopoly of US based database vendors with innovative technologies.

Big Data as an Opportunity for German Technology Providers. A multitude of German vendors, research institutions and universities are well positioned. In addition to SAP's HANA in-memory technology and Software AG's Terra Cotta system, small and medium sized enterprises such as Exasol or ParSteam as well as innovative technologies from the university research community such as Stratosphere, Hyper, HAIL/Hadoop++ which could be commercialized have to be mentioned. These companies that stem from university cooperations, have already won various international awards and beat the US-American competitors in crucial benchmarks like the TPV Benchmark.

Innovative Research Prototypes at German Universities. German universities additional innovative systems and prototypes such as Stratosphere (TU Berlin, HU Berlin, Hasso-Plattner Institute), Hyper (TU Munich) Hadapt/HAIL (University Saarland) with new, disruptive technologies in the field of efficient specification and scalable execution of machine learning methods and mixes workloads have been created.

Data Marketplaces

The high technical, organizational and personnel efforts necessary for the preparation and execution of reliable and comprehensible big data analyses are still an obstacle for many companies. Data marketplaces provide extracted and integrated data in a centralized manner and can thus lower the cost for individual companies, which can obtain the cleansed and integrated data, significantly. The central entry point of a data market place eases the access to such services and data. Furthermore the data marketplace serves as a data integration platform across customers and merchants, in particular for the collective storage, analyses and re-use of data. Information marketplaces enable in particular small and medium sized enterprises to analyse such data and to apply the results to even out their competitive disadvantage

7 Recommendations for Decision-makers

Big data is a general concept embracing technologies to collect, process and present large, heterogeneous quantities of data which accrue in very short periods and can be used for very rapid decisions. Big data can thus foster disruptive changes on markets and in companies. These disruptive changes can result in substantial opportunities and competitive advantages for businesses in Germany. At the same time, big data entails risks. In a globalised and closely integrated economy, it is necessary to shape a policy environment which permits German firms to utilise the opportunities of big data while effectively controlling the related risks. The following section presents three premises and six recommendations for action for the development of an effective environment for big data. The premises stress central aspects of promoting big data in Germany which serve as a basis for all the recommended actions. The recommended actions represent important orientations for developments which make it possible to utilise the opportunities of big data and to control the related risks.

Premise 1: Education and Management of Expectations. The competitive advantages offered by big data necessitate an objective discussion of the opportunities and risks of big data technologies and their use, with broad participation from commerce, government, society and academia.

Premise 2: Responsible Use of Data. The competitive advantages offered by big data require clear rules preconditions and limitations to the responsible use of personal data.

Premise 3: Small and Medium Sized Firms are an Important Target Group. The competitive advantages of big data require the targeted support of small and medium-sized providers and users of big data technology.

Recommended Action 1: Make Use of Currently Unused Data to Optimise Operative Business Processes. The compilation and preparation of data is resource-intensive and cost-intensive. Companies are therefore reluctant to invest in big data. It is therefore necessary to support pilot projects which help companies to better assess the cost and benefit of big data. Data archives from manufacturing, development or operations are particularly suited to this. These do not generally contain much personal data, so that the legal barriers are lower. Furthermore, they can be allocated to specific operative process optimisations, and this makes it easier to understand the benefits of using the technology. If sensible commercial use can be made of the potential of big data, it will be possible to derive robust arguments in favour of using big data.

Recommended Action 2: Building and Strengthening Ecosystems for Data Services. Big data establishes the technical framework for data services, i.e. data and data analyses become commercial assets. Support should therefore be given to measures which provide large quantities of data for analysis by third parties as well. In particular, companies should be enabled to trade, exchange or disclose data and data analyses. These measures should make it possible to make robust statements about the structure, development and sale of such data services. In view of the novel nature and volatility of such a market, the emergence of complementary data providers for Germany's core sectors like industry, healthcare and mobility should be supported for important public data.

Recommended Action 3: Strengthening German Technology Providers for Big Data (Technology Push). In the field of big data technologies, Germany is very well positioned thanks to research at universities and developments in companies. For this reason, measures should be supported which aim to commercialise these technologies in Germany and internationally. This can particularly take place via close co-operation between technology providers and potential users. This gives potential users the possibility to intervene in their own interest in the final phases of technology development; the provider can orient its product or service better to the specific needs of the user. The aim is to boost the prospects of success of the German big data technology providers entering the market (technology push).

Recommended Action 4: Establishment and Strengthening of Sector Specific and Cross Sector Innovation Networks for Big Data (Market Pull). One major challenge for the broad use of the potential of big data is to identify or establish the demand amongst potential users (market pull). Support should be given to measures which enable potential users and technology providers of big data to join forces with innovation networks and to develop data-driven innovations. Here, the form of the innovation networks ensures the sustainability of the innovation beyond the individual company and creates the possibility for new forms of co-operation on the use of data. The main emphasis here is on commercial aspects and the realisation of big data potential in new products, services and business models. If it proves possible to establish specific and robust requirements for big data technology here, there will be opportunities for companies to develop corresponding products and services.

Recommended Action 5: Boosting the Legal Certainty in the Use of Big Data and Overcoming Existing Barriers. It is already possible to use big data in compliance with the law. Despite this, the current legal situation makes companies reluctant to make effective use of the full commercial potential of big data applications. Adapting the legal framework to the current state of the art, particularly in the field of data protection law and copyright law, could make a decisive contribution towards overcoming barriers and increasing legal certainty.

Recommended Action 6: Expansion of Training Courses for Data Science as a Key Skill. There is an urgent need for training courses for the quantitative and qualitative analysis of large heterogeneous data quantities with low latency. Here, support should be given to products and services which integrate the systems view of big data with the analytical view and with a responsible, legally secure use of big data. Similarly the economic aspects of big data must be taken into account.

Colophon

Editors
Prof. Volker Markl, TU-Berlin
Prof. Thomas Hoeren, WWU-Münster
Prof. Helmut Krcmar, TU-München

Authors
This executive summary is a translated excerpt of a detailed report commissioned by the Federal Ministry for Economic Affairs and Energy. The German title of the detailed report is "Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen (Big Data Management)" The authors of the detailed report are: Volker Markl, Alexander Löser, Thomas Hoeren, Helmut Krcmar, Holmer Hemsen, Michael Schermann, Matthias Gottlieb, Christoph Buchmüller, Philip Uecker, Till Bitter

Photos/Illustrations
Photo (S. 3) Holmer Hemsen/TU-Berlin
Illustrations (S. 9-11) Own work based on the outcome of a questionnaire based study made for the big data management report

Editoral department
TU-Berlin/FG DIMA

Translation
Christoph Boden, Johannes Kirschnick, Juan Soto (TU-Berlin), kindly supported by the translation service of the Federal Ministry for Economic Affairs and Energy

Publication date
March, 2014; Version 1.0