Chapter 8
Big Data Usage

Tilman Becker

8.1 Introduction

One of the core business tasks of advanced data usage is the support of business decisions. Data usage is a wide field that is addressed in this chapter by viewing data usage from various perspectives, including the underlying technology stacks, trends in various sectors, the impact on business models, and requirements on human–computer interaction.

The full life-cycle of information is covered in this book, with previous chapters covering data acquisition, storage, analysis, and curation. The position of big data usage within the overall big data value chain can be seen in Fig. 8.1. Data usage covers the business goals that need access to such data, its analyses, and the tools needed to integrate the analyses in business decision-making.

The process of decision-making includes reporting, exploration of data (browsing and lookup), and exploratory search (finding correlations, comparisons, what-if scenarios, etc.). The business value of such information logistics is twofold: (1) control over the value chain and (2) transparency of the value chain. The former is generally independent from big data; the latter, however, provides opportunities and requirements for data markets and services.

Big data influences the validity of data-driven decision-making in the future. Influencing factors are (1) the time range for decisions/recommendations, from short term to long term and (2) the various databases (in a non-technical sense) from past, historical data to current and up-to-date data.

New data-driven applications will strongly influence the development of new markets. A potential blocker of such developments is always the need for new

T. Becker (✉)
German Research Centre for Artificial Intelligence (DFKI), Stuhlsatzenhausweg 3, 66123
Saarbrücken, Germany
e-mail: tilman.becker@dfki.de

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_8

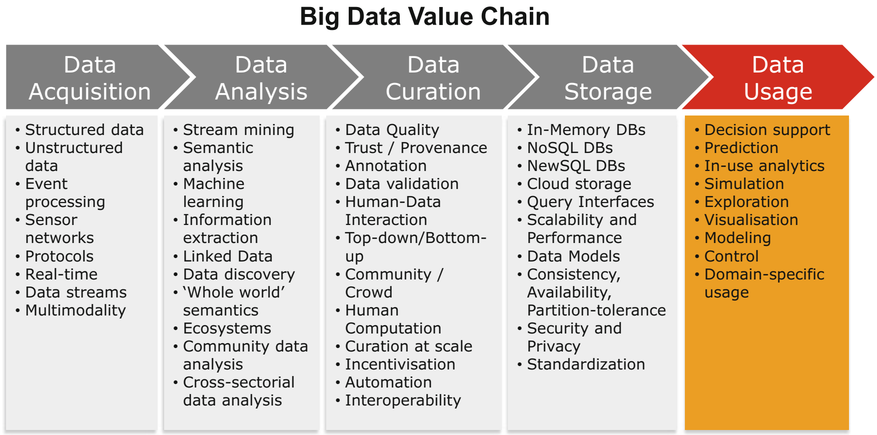

Fig. 8.1 Data usage in the big data value chain

partner networks (combination of currently separate capabilities), business processes, and markets.

A special area of use cases for big data is the manufacturing, transportation, and logistics sector. These sectors are undergoing a transformational change as part of an industry-wide trend, called "Industry 4.0", which originates in the digitization and interlinking of products, production facilities, and transportation infrastructure as part of the developing "Internet of Things". Data usage has a profound impact in these sectors, e.g. applications of predictive analysis in maintenance are leading to new business models as the manufacturers of machinery are in the best position to provide big data-based maintenance. The emergence of cyber-physical systems (CPS) for production, transportation, logistics, and other sectors brings new challenges for simulation and planning, for monitoring, control, and interaction (by experts and non-experts) with machinery or data usage applications.

On a larger scale, new services and a new service infrastructure is required. Under the title "smart data" and smart data services, requirements for data and also service markets are formulated. Besides the technology infrastructure for the interaction and collaboration of services from multiple sources, there are legal and regulatory issues that need to be addressed. A suitable service infrastructure is also an opportunity for SMEs to take part in big data usage scenarios by offering specific services, e.g., through data usage service marketplaces.

Access to data usage is given through specific tools and in turn through query and scripting languages that typically depend on the underlying data stores, their execution engines, APIs, and programming models. In Sect. 8.5.1, different technology stacks and some of the trade-offs involved are discussed. Section 8.5.2 presents general aspects of decision support, followed by a discussion of specific access to analysis results through visualization and new explorative interfaces. Human–computer interaction will play a growing role in decision support since many cases cannot rely on pre-existing models of correlation. In such cases, user interfaces (e.g. in data visualization for visual analytics) must support an exploration of the data and their potential connections. Emerging trends and future requirements are presented in Sect. 8.6 with special emphasis on Industry 4.0 and the emerging need for smart data and smart services.

8.2 Key Insights for Big Data Usage

The key insights for big data usage identified are as follows:

Predictive Analytics A prime example for the application of predictive analytics is in predictive maintenance based on sensor and context data to predict deviations from standard maintenance intervals. Where data points to a stable system, maintenance intervals can be extended, leading to lower maintenance costs. Where data points to problems before reaching a scheduled maintenance, savings can be even higher if a breakdown, repair cost, and downtimes can be avoided. Information sources go beyond sensor data and tend to include environmental and context data, including usage information (e.g. high load) of the machinery. As predictive analysis depends on new sensors and data processing infrastructure, large manufacturers are switching their business model and investing in new infrastructure themselves (realizing scale effects on the way) and leasing machinery to their customers.

Industry 4.0 A growing trend in manufacturing is the employment of cyber-physical systems. It brings about an evolution of old manufacturing processes, on the one hand making available a massive amount of sensor and other data and on the other hand bringing the need to connect all available data through communication networks and usage scenarios that reap the potential benefits. Industry 4.0 stands for the entry of IT into the manufacturing industry and brings with it a number of challenges for IT support. This includes services for diverse tasks such as planning and simulation, monitoring and control, interactive use of machinery, logistics and enterprise resource planning (ERP), predictive analysis, and eventually prescriptive analysis where decision processes can be automatically controlled by data analysis.

Smart Data and Service Integration When further developing the scenario for Industry 4.0 above, services that solve the tasks at hand come into focus. To enable the application of smart services to deal with the big data usage problems, there are technical and organizational matters. Data protection and privacy issues, regulatory issues, and new legal challenges (e.g. with respect to ownership issues for derived data) must all be addressed.

On a technical level, there are multiple dimensions along which the interaction of services must be enabled: on a hardware level from individual machines, to facilities, to networks; on a conceptual level from intelligent devices to intelligent systems and decisions; on an infrastructure level from IaaS to PaaS and SaaS to new services for big data usage and even to business processes and knowledge as a service.

Interactive Exploration When working with large volumes of data in large variety, the underlying models for functional relations are oftentimes missing. This means data analysts have a greater need for exploring datasets and analyses. This is addressed through visual analytics and new and dynamic ways of data visualization, but new user interfaces with new capabilities for the exploration of data are needed. Integrated data usage environments provide support, e.g., through history mechanisms and the ability to compare different analyses, different parameter settings, and competing models.

8.3 Social and Economic Impact for Big Data Usage

One of the most important impacts of big data usage scenarios is the discovery of new relations and dependencies in the data that lead, on the surface, to economic opportunities and more efficiency. On a deeper level, big data usage can provide a better understanding of these dependencies, making the system more transparent and supporting economic as well as social decision-making processes (Manyika et al. 2011). Wherever data is publicly available, social decision-making is supported; where relevant data is available on an individual-level, personal decision-making is supported. The potential for transparency through big data usage comes with a number of requirements: (1) regulations and agreements on data access, ownership, protection, and privacy, (2) demands on data quality (e.g. on the completeness, accuracy, and timeliness of data), and (3) access to the raw data as well as access to appropriate tools or services for big data usage.

Transparency thus has an economic and social and personal dimension. Where the requirements listed above can be met, decisions become transparent and can be made in a more objective, reproducible manner, where the decision processes are open to involve further players.

The current economic drivers of big data usage are large companies with access to complete infrastructures. These include sectors like advertising at Internet companies and sensor data from large infrastructures (e.g. smart grids or smart cities) or for complex machinery (e.g. airplane engines). In the latter examples, there is a trend towards even closer integration of data usage at large companies as the big data capabilities remain with the manufactures (and not the customers), e.g. when engines are only rented and the big data infrastructure is owned and managed by the manufacturers.

There is a growing requirement for standards and accessible markets for data as well as for services to manage, analyse, and exploit further uses of data. Where such requirements are met, opportunities are created for SMEs to participate in more complex use cases for big data usage. Section 8.5.2.1 discusses these requirements for smart data and corresponding smart data services.

8.4 Big Data Usage State-of-the-Art

This section provides an overview of the current state of the art in big data usage, addressing briefly the main aspects of the technology stacks employed and the subfields of decision support, predictive analysis, simulation, exploration, visualization, and more technical aspects of data stream processing. Future requirements and emerging trends related to big data usage will be addressed in Sect. 8.6.

8.4.1 Big Data Usage Technology Stacks

Big data applications rely on the complete data value chain that is covered in the BIG project, starting at data acquisition, including curation, storage, analysis, and being joined for data usage. On the technology side, a big data usage application relies on a whole stack of technologies that cover the range from data stores and their access to processing execution engines that are used by query interfaces and languages.

It should be stressed that the complete big data technology stack can be seen as much broader, i.e., encompassing the hardware infrastructure, such as storage systems, servers, datacentre networking infrastructure, corresponding data organization and management software, as well as a whole range of services ranging from consulting and outsourcing to support and training on the business side as well as the technology side.

Actual user access to data usage is given through specific tools and in turn through query and scripting languages that typically depend on the underlying data stores, their execution engines, APIs, and programming models. Some examples include SQL for classical relational database management systems (RDBMS), Dremel and Sawzall for Google's file system (GFS), and MapReduce, Hive, Pig, and Jaql for Hadoop-based approaches, Scope for Microsoft's Dryad and CosmosFS, and many other offerings, e.g. Stratosphere's[1] Meteor/Sopremo and ASTERIX's AQL/Algebricks.

Analytics tools that are relevant for data usage include SystemT (IBM, for data mining and information extraction) and Matlab (U. Auckland and Mathworks, resp. for mathematical and statistical analysis), tools for business intelligence and analytics (SAS Analytics (SAS), Vertica (HP), SPSS (IBM)), tools for search and indexing (Lucene and Solr (Apache)), and specific tools for visualization (Tableau, Tableau Software). Each of these tools has its specific area of application and covers different aspects of big data.

The tools for big data usage support business activities that can be grouped into three categories: lookup, learning, and investigating. The boundaries are sometimes fuzzy and learning and investigating might be grouped as examples of exploratory search. Decision support needs access to data in many ways, and as big data more

[1] Stratosphere is further developed in the Apache Flink project.

often allows the detection of previously unknown correlations, data access must be more often from interfaces that enable exploratory search and not mere access to predefined reports.

8.4.1.1 Trade-Offs in Big Data Usage Technologies

An in-depth case study analysis of a complete big data application was performed to determine the decisions involved in weighing the advantages and disadvantages of the various available components of a big data technology stack. Figure 8.2 shows the infrastructure used for Google's YouTube Data Warehouse (YTDW) as detailed in Chattopadhyay (2011). Some of the core lessons learned by the YouTube team include an acceptable trade-off in functionality when giving priority to low-latency queries. This justified the decision to stick with the ([Dremel tool (for querying large datasets) that has acceptable drawbacks in expressive power (when compared to SQL-based tools), yet provides low-latency results and scales to what Google considers "medium" scales. Note, however, that Google is using "trillions of rows in seconds", and running on "thousands of CPUs and petabytes of data", processing "quadrillions of records per month". While Google regards this as medium scale, this might be sufficient for many applications that are clearly in the realms of big data. Table 8.1 shows a comparison of various data usage technology components used in the YTDW, where latency refers to the time the systems need to answer request; scalability to the ease of using ever larger datasets; SQL refers to the (often preferred) ability to use SQL (or similar) queries; and power refers to the expressive power of search queries.

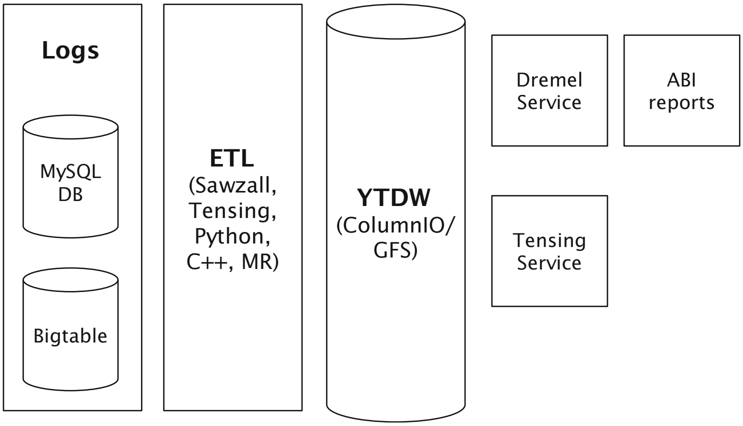

Fig. 8.2 The YouTube Data Warehouse (YTDW) infrastructure. Derived from Chattopadhyay (2011)

Table 8.1 Comparison of data usage technologies used in YTDW. Source: Chattopadhyay (2011)

	Sawzall	Tenzing	Dremel
Latency	High	Medium	Low
Scalability	High	High	Medium
SQL	None	High	Medium
Power	High	Medium	Low

8.4.2 Decision Support

Current decision support systems—as far as they rely on static reports—use these techniques but do not allow sufficient dynamic usage to reap the full potential of exploratory search. However, in increasing order of complexity, these groups encompass the following business goals:

- **Lookup:** On the lowest level of complexity, data is merely retrieved for various purposes. These include fact retrieval and searches for known items, e.g. for verification purposes. Additional functionalities include navigation through datasets and transactions.
- **Learning:** On the next level, these functionalities can support knowledge acquisition and interpretation of data, enabling comprehension. Supporting functionalities include comparison, aggregation, and integration of data. Additional components might support social functions for data exchange. Examples for learning include simple searches for a particular item (knowledge acquisition), e.g. a celebrity and their use in advertising (retail). A big data search application would be expected to find all related data and present an integrated view.
- **Investigation:** On the highest level of decision support systems, data can be analysed, accreted, and synthesized. This includes tool support for exclusion, negation, and evaluation. At this level of analysis, true discoveries are supported and the tools influence planning and forecasting. Higher levels of investigation (discovery) will attempt to find important correlations, say the influence of seasons and/or weather on sales of specific products at specific events. More examples, in particular of big data usage for high-level strategic business decisions, are given in Sect. 8.6 on future requirements.

At an even higher level, these functionalities might be (partially) automated to provide predictive and even normative analyses. The latter refers to automatically derived and implemented decisions based on the results of automatic (or manual) analysis. However, such functions are beyond the scope of typical decision support systems and are more likely to be included in complex event processing (CEP) environments where the low latency of automated decision is weighed higher than the additional safety of a human-in-the-loop that is provided by *decision support systems*.

8.4.3 Predictive Analysis

A prime example of predictive analysis is predictive maintenance based on big data usage. Maintenance intervals are typically determined as a balance between a costly, high frequency of maintenance and an equally costly danger of failure before maintenance. Depending on the application scenario, safety issues often mandate frequent maintenance, e.g., in the aerospace industry. However, in other cases the cost of machine failures is not catastrophic and determining maintenance intervals becomes a purely economic exercise.

The assumption underlying predictive analysis is that given sufficient sensor information from a specific machine and a sufficiently large database of sensor and failure data from this machine or the general machine type, the specific time to failure of the machine can be predicted more accurately. This approach promises to lower costs due to:

- *Longer maintenance intervals* as "unnecessary" interruptions of production (or employment) can be avoided when the regular time for maintenance is reached. A predictive model allows for an extension of the maintenance interval, based on current sensor data.
- *Lower number of failures* as the number of failures occurring earlier than scheduled maintenance can be reduced based on sensor data and predictive maintenance calling for earlier maintenance work.
- *Lower costs for failures* as potential failures can be predicted by predictive maintenance with a certain advance warning time, allowing for scheduling maintenance/exchange work, lowering outage times.

8.4.3.1 New Business Model

The application of predictive analytics requires the availability of sensor data for a specific machine (where "machine" is used as a fairly generic term) as well as a comprehensive dataset of sensor data combined with failure data.

Equipping existing machinery with additional sensors, adding communication pathways from sensors to the predictive maintenance services, etc., can be a costly proposition. Based on experiencing reluctance from their customers in such investments, a number of companies (mainly manufacturers of machines) have developed new business models addressing these issues.

Prime examples are GE wind turbines and Rolls Royce airplane engines. Rolls Royce engines are increasingly offered for rent, with full-service contracts including maintenance, allowing the manufacturer to lift the benefits from applying predictive maintenance. By correlating the operational context with engine sensor data, failures can be predicted early, reducing (the costs of) replacements, allowing for planned maintenance rather than just scheduled maintenance. GE OnPoint solutions offer similar service packages that are sold in conjunction with GE engines.[2]

8.4.4 Exploration

Exploring big datasets and the corresponding analytics results can be distributed across multiple sources and formats (e.g. new portals, travel blogs, social networks, web services, etc.). To answer complex questions—e.g. “Which astronauts have been on the moon?”, “Where is the next Italian restaurant with high ratings?”, “Which sights should I visit in what order?”—users have to start multiple requests to multiple, heterogeneous sources and media. Finally, the results have to be combined manually.

Support for the human trial-and-error approach can add value by providing intelligent methods for automatic information extraction and aggregation to answer complex questions. Such methods can transform the data analysis process to become explorative and iterative. In a first phase, relevant data is identified and then a second learning phase context is added for such data. A third exploration phase allows various operations for deriving decisions from the data or transforming and enriching the data.

Given the new complexity of data and data analysis available for exploration, there are a number of emerging trends in explorative interfaces that are discussed in Sect. 8.5.2.4 on complex exploration.

8.4.5 Iterative Analysis

An efficient, parallel processing of iterative data streams brings a number of technical challenges. Iterative data analysis processes typically compute analysis results in a sequence of steps. In every step, a new intermediate result or state is computed and updated. Given the high volumes in big data applications, computations are executed in parallel, distributing, storing, and managing the state efficiently across multiple machines. Many algorithms need a high number of iterations to compute the final results, requiring low latency iterations to minimize overall response times. However, in some applications, the computational effort is reduced significantly between the first and the last iterations. Batch-based systems such as Map/Reduce (Dean and Ghemawat 2008) and Spark (Apache 2014) repeat all computations in every iteration even when the (partial) results do not change.

[2] See http://www.aviationpros.com/press_release/11239012/tui-orders-additional-genx-powered-boeing-787s

Truly iterative dataflow systems like Stratosphere (Stratosphere 2014) of specialized graph systems like GraphLab (Low et al. 2012) and Google Pregel (Malewicz et al. 2010) exploit such properties and reduce the computational cost in every iteration.

Future requirements on technologies and their applications in big data usage are described in Sect. 8.5.1.3, covering aspects of pipelines versus materialization and error tolerance.

8.4.6 Visualization

Visualizing the results of an analysis including a presentation of trends and other predictions by adequate visualization tools is an important aspect of big data usage. The selection of relevant parameters, subsets, and features is a crucial element of data mining and machine learning with many cycles needed for testing various settings. As the settings are evaluated on the basis of the presented analysis results, a high-quality visualization allows for a fast and precise evaluation of the quality of results, e.g., in validating the predictive quality of a model by comparing the results against a test dataset. Without supportive visualization, this can be a costly and slow process, making visualization an important factor in data analysis.

For using the results of data analytics in later steps of a data usage scenario, for example, allowing data scientists and business decision-makers to draw conclusions from the analysis, a well-selected visual presentation can be crucial for making large result sets manageable and effective. Depending on the complexity of the visualizations, they can be computationally costly and hinder interactive usage of the visualization.

However, explorative search in analytics results is essential for many cases of big data usage. In some cases, the results of a big data analysis will be applied only to a single instance, say an airplane engine. In many cases, though, the analysis dataset will be as complex as the underlying data, reaching the limits of classical statistical visualization techniques and requiring interactive exploration and analysis (Spence 2006; Ward et al. 2010). In Shneiderman's seminal work on visualization (Shneiderman 1996), he identifies seven types of tasks: overview, zoom, filter, details-on-demand, relate, history, and extract.

Yet another area of visualization applies to data models that are used in many machine-learning algorithms and differ from traditional data mining and reporting applications. Where such data models are used for classification, clustering, recommendations, and predictions, their quality is tested with well-understood datasets. Visualization supports such validation and the configuration of the models and their parameters.

Finally, the sheer size of datasets is a continuous challenge for visualization tools that is driven by technological advances in GPUs, displays, and the slow adoption of immersive visualization environments such as caves, VR, and AR. These aspects are covered in the fields of scientific and information visualization.

The following section elaborates the application of visualization for big data usage, known as visual analytics. Section 8.5.1.4 presents a number of research challenges related to visualization in general.

8.4.6.1 Visual Analytics

A definition of visual analytics, taken from Keim et al. (2010) recalls first mentions of the term in 2004. More recently, the term is used in a wider context, describing a new multidisciplinary field that combines various research areas including visualisation, human–computer interaction, data analysis, data management, geo-spatial and temporal data processing, spatial decision support and statistics.

The "Vs" of big data affect visual analytics in a number of ways. The **volume** of big data creates the need to visualize high dimensional data and their analyses and to display multiple data types such as linked graphs. In many cases interactive visualization and analysis environments are needed that include dynamically linked visualizations. Data **velocity** and the dynamic nature of big data calls for correspondingly dynamic visualizations that are updated much more often than previous, static reporting tools. Data **variety** presents new challenges for cockpits and dashboards.

The main new aspects and trends are:

- Interactivity, visual queries, (visual) exploration, multi-modal interaction (touchscreen, input devices, AR/VR)
- Animations
- User adaptivity (personalization)
- Semi-automation and alerting, CEP (complex event processing), and BRE (business rule engines)
- Large variety in data types, including graphs, animations, microcharts (Tufte), gauges (cockpit-like)
- Spatiotemporal datasets and big data applications addressing geographic information systems (GIS)
- Near real-time visualization. Sectors finance industry (trading), manufacturing (dashboards), oil/gas—CEP, BAM (business activity monitoring)
- Data granularity varies widely
- Semantics

Use cases for visual analytics include multiple sectors, e.g. marketing, manufacturing, healthcare, media, energy, transportation (see also the use cases in Sect. 8.6), but also additional market segments such as software engineering.

A special case of visual analytics that is spearheaded by the US intelligence community is visualization for cyber security. Due to the nature of this market segment, details can be difficult to obtain; however there are publications available, e.g. the VizSec conferences.[3]

8.5 Future Requirements and Emerging Trends for Big Data Usage

This section provides an overview of future requirements and emerging trends that resulted from the task force's research.

8.5.1 Future Requirements for Big Data Usage

As big data usage is becoming more important, there are issues on the underlying assumptions that become more important. The key issue is a necessary validation of the underlying data. The following quote as attributed to Ronald Coase, winner of the Nobel Prize in economics in 1991, put it as a joke alluding to the inquisition: "If you torture the data long enough, it [they] will confess to anything".

On a more serious note there are some common misconceptions in big data usage:

1. Ignoring modelling and instead relying on correlation rather than an understanding of causation.
2. The assumption that with enough—or even all (see next point)—data available, no models are needed (Anderson 2008).
3. Sample bias. Implicit in big data is the expectation that *all* data will (eventually) be sampled. This is rarely ever true; data acquisition depends on technical, economical, and social influences that create sample bias.
4. Overestimation of accuracy of analysis: it is easy to ignore false positives.

To address these issues, the following future requirements will gain importance:

1. Include more modelling, resort to simulations, and correct (see next point) for sample bias.
2. Understand the data sources and the sample bias that is introduced by the context of data acquisition. Create a model of the real, total dataset to correct for sample bias.
3. Data and analysis transparency: If the data and the applied analyses are known, it is possible to judge what the (statistical) chances are that correlations are not

[3] http://www.vizsec.org

only "statistically significant" but also that the number of tested, possible correlations is not big enough to make the finding of some correlation almost inevitable.

With these general caveats as background, the key areas that are expected to govern the future of big data usage have been identified:

- Data quality in big data usage
- Tool performance
- Strategic business decisions
- Human resources, big data specific positions

The last point is exemplified by a report on the UK job market in big data (e-skills 2013) where demand is growing strongly. In particular, the increasing number of administrators sought shows that big data is growing from experimental status to a core business unit.

8.5.1.1 Specific Requirements

Some general trends are already identifiable and can be grouped into the following requirements:

- Use of big data for marketing purposes
- Detect abnormal events of incoming data in real time
- Use of big data to improve efficiency (and effectiveness) in core operations
  - Realizing savings during operations through real-time data availability, more fine-grained data, and automated processing
  - Better data basis for planning of operational details and new business processes
  - Transparency for internal and external (customers) purposes
- Customization, situation adaptivity, context-awareness, and personalization
- Integration with additional datasets
  - Open data
  - Data obtained through sharing and data marketplaces
- Data quality issues where data is not curated or provided under pressure, e.g., to acquire an account in a social network where the intended usage is anonymous
- Privacy and confidentiality issues, data access control
- Interfaces
  - Interactive and flexible, ad hoc analyses to provide situation-adaptive and context-aware reactions, e.g. recommendations
  - Suitable interfaces to provide access to big data usage in non-office environments, e.g. mobile situations, factory floors, etc.
  - Tools for visualization, query building, etc.

- Discrepancy between the technical know-how necessary to execute data analysis (technical staff) and usage in business decisions (by non-technical staff)
- Need for tools that enable early adoption. As the developments in industry are perceived to be accelerating, the head start from early adoption is also perceived as being of growing importance and a growing competitive advantage.

8.5.1.2 Industry 4.0

For applications of big data in areas such as manufacturing, energy, transportation, and even health, wherever intelligent machines are involved in the business process, there is a need for aligning hardware technology (i.e. machines and sensors) with software technology (i.e. the data representation, communication, storage, analysis, and control of the machinery). Future developments in embedded systems that are developing into "cyber-physical systems" will need to synchronize the joint development of hardware (computing, sensing, and networking) and software (data formats, operating systems, and analysis and control systems).

Industrial suppliers are beginning to address these issues. GE software identifies "However well-developed industrial technology may be, these short-term and long-term imperatives cannot be realized using today's technology alone. The software and hardware in today's industrial machines are very interdependent and closely coupled, making it hard to upgrade software without upgrading hardware, and vice versa" (Chauhan 2013).

On the one hand this adds a new dependency to big data usage, namely the dependency on hardware systems and their development and restrictions. On the other hand, it opens new opportunities to address more integrated systems with big data usage applications at the core of supporting business decisions.

8.5.1.3 Iterative Data Streams

There are two prominent areas of requirements for efficient and robust implementations of big data usage that relate to the underlying architectures and technologies in distributed, low-latency processing of large datasets and large data streams.

- **Pipelining and materialization:** High data rates pose a special challenge for data stream processing. The underlying architectures are based on a pipeline approach where processed data can be handed to the next processing step with very low delay to avoid pipeline congestion. In cases where such algorithms do not exist, data is collected and stored before being processed. Such approaches are called "materialization". Low latency for queries can typically only be realized in pipelining approaches.
- **Error tolerance:** Fault tolerance and error minimization are an important challenge for pipelining systems. Failures in compute nodes are common and

can cause parts of the analysis result to be lost. Parallel systems must be designed in a robust way to overcome such faults without failing. A common approach are continuous *check points* at which intermediate results are saved, allowing the reconstruction of a previous state in case of an error. Saving data at checkpoints is easy to implement, yet results in high execution costs due to the synchronization needs and storage costs when saving to persistent storage. New alternative algorithms use optimistic approaches that can recreate valid states allowing the continuation of computing. Such approaches add costs only in cases of errors but are applicable only in restricted cases.

8.5.1.4 Visualization

There are a number of future trends that need to be addressed in the area of visualization and visual analytics in the medium to far future, for example (Keim et al. 2010):

- Visual perception and cognitive aspects
- “Design” (visual arts)
- Data quality, missing data, data provenance
- Multi-party collaboration, e.g., in emergency scenarios
- Mass-market, end user visual analytics

In addition, Markl et al. (2013) compiled a long list of research questions from which the following are of particular importance to data usage and visualization:

- How can visualization support the process of constructing data models for prediction and classification?
- Which visualization technologies can support an analyst in explorative analysis?
- How can audio and video (animations) be automatically collected and generated for visual analytics?
- How can meta-information such as semantics, data quality, and provenance be included into the visualization process?

8.5.2 Emerging Paradigms for Big Data Usage

A number of emerging paradigms for big data usage have been identified that fall into two categories. The first category encompasses all aspects of integration of big data usage into larger business processes and the evolution towards a new trend called “smart data”. The second trend is much more local and concerns the interface tools for working with big data. New exploration tools will allow data scientists and analysts in general to access more data more quickly and support decision-making by finding trends and correlations in the dataset that can be grounded in models of the underlying business processes.

There are a number of technology trends that are emerging (e.g. in-memory databases) that allow for a sufficiently fast analysis to enable explorative data analysis and decision support. At the same time, new services are developing, providing data analytics, integration, and transformation of big data to organizational knowledge.

As in all new digital markets, the development is driven in part by start-ups that fill new technology niches; however, the dominance of big players is particularly important as they have much easier access to big data. The transfer of technology to SMEs is faster than in previous digital revolutions; however, appropriate business cases for SMEs are not easy to design in isolation and typically involve the integration into larger networks or markets.

8.5.2.1 Smart Data

The concept of smart data is defined as the effective application of big data that is successful in bringing measurable benefits and has a clear meaning (semantics), measurable data quality, and security (including data privacy standards).[4]

Smart data scenarios are thus a natural extension of big data usage in any economically viable context. These can be new business models that are made possible by innovative applications of data analysis, or improving the efficiency/ profitability of existing business models. The latter are easy to start with as data is available and, as it is embedded in existing business processes, already has an assigned meaning (semantics) and business structure. Thus, it is the added value of guaranteed data quality and existing metadata that can make big data usage become a case of smart data.

Beyond the technical challenges, the advent of smart data brings additional challenges:

1. Solving regulatory issues regarding data ownership and data privacy (Bitkom 2012).
2. Making data more accessible by structuring through the addition of metadata, allowing for the integration of separate data silos (Bertolucci 2013).
3. Lifting the benefits from already available open data and linked data sources. Their market potential is currently not fully realized (Groves et al. 2013).

The main potential of data usage, according to Lo (2012), is found in the optimization of business processes, improved risk management, and market-oriented product development. The purpose of enhanced big data usage as smart data is in solving social and economical challenges in many sectors, including energy, manufacturing, health, and media.

[4] This section reflects the introduction of smart data as stated in a broadly supported memorandum, available at http://smart-data.fzi.de/memorandum/

For SMEs, the focus is on the integration into larger value chains that allow multiple companies to collaborate to give SMEs access to the effects of scale that underlie the promise of big data usage. Developing such collaborations is enabled by smart data when the meaning of data is explicit, allowing for the combination of planning, control, production, and state information data beyond the limits of each partnering company.

Smart data creates requirements in four areas: semantics, data quality, data security and privacy, and metadata.

Semantics Understanding and having available the meaning of datasets enables important steps in smart data processing:

- Interoperability
- Intelligent processing
- Data integration
- Adaptive data analysis

Metadata As a means to encode and store the meaning (semantics) of data. Metadata can also be used to store further information about data quality, provenance, usage rights, etc. Currently there are many proposals but no established standards for metadata.

Data Quality The quality and provenance of data is one of the well-understood requirements for big data (related to one of the “Vs”, i.e. “veracity”).

Data Security and Privacy These separate, yet related, issues are particularly influenced by existing regulatory standards. Violations of data privacy laws can easily arise from processing of personal data, e.g. movement profiles, health data, etc. Although such data can be enormously beneficial, violations of data privacy laws carry severe punishments. Other than doing away with such regulations, methods for anonymization (ICO 2012) and pseudonymization (Gowing and Nickson 2010) can be developed and used to address these issues.

8.5.2.2 Big Data Usage in an Integrated and Service-Based Environment

The continuing integration of digital services (Internet of Services), smart digital products (Internet of things), and production environments (Internet of Things, Industry 4.0) includes the usage of big data in most integration steps. A recent study by General Electric examined the various dimensions of integration within the airline industry (Evans and Annunziata 2012). Smart products like a turbine are integrated into larger machines, and in the first example this is an airplane. Planes are in turn part of whole fleets that operate in a complex network of airports, maintenance hangars, etc. At each step, the current integration of the business processes is extended by big data integration. The benefits for optimization can be harvested at each level (assets, facility, fleets, and the entire network) and by integrating knowledge from data across all steps.

8.5.2.3 Service Integration

The infrastructure within which big data usage will be applied will adapt to this integration tendency. Hardware and software will be offered as services, all integrated to support big data usage. See Fig. 8.3 for a concrete picture of the stack of services that will provide the environment for "Beyond technical standards and protocols, new platforms that enable firms to build specific applications upon a shared framework/architecture [are necessary]", as foreseen by the GE study or the "There is also a need for on-going innovation in technologies and techniques that will help individuals and organisations to integrate, analyse, visualise, and consume the growing torrent of big data", as sketched by McKinsey's study (Manyika et al. 2011).

Figure 8.3 shows big data as part of a virtualized service infrastructure. At the bottom level, current hardware infrastructure will be virtualized with cloud computing technologies; hardware infrastructure as well as platforms will be provided as services. On top of this cloud-based infrastructure, software as a service (SaaS) and on top of this business processes as a service (BPaaS) can be built. In parallel, big data will be offered as a service and embedded as the precondition for knowledge services, e.g. the integration of semantic technologies for analysis of unstructured and aggregated data. Note that big data as a service may be seen as extending a layer between PaaS and SaaS.

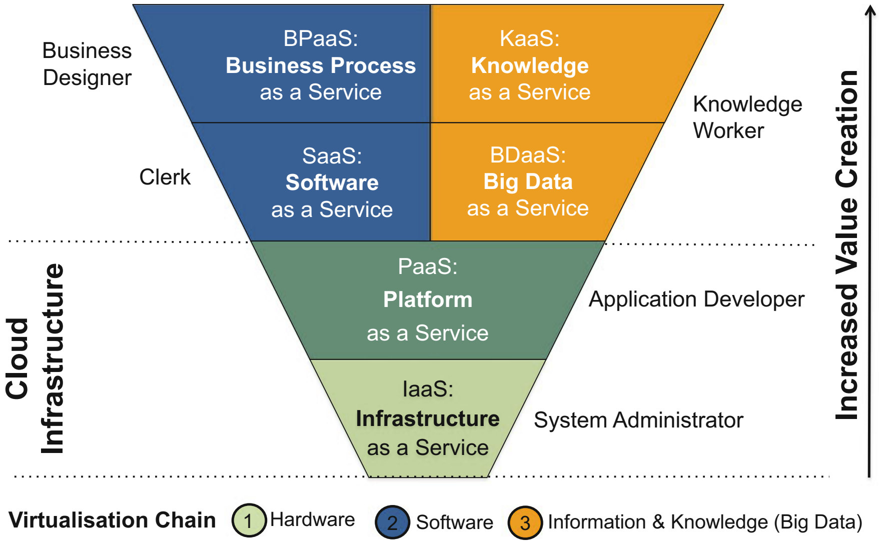
Business Designer
BPaaS: Business Process as a Service
KaaS: Knowledge as a Service
Knowledge Worker
Clerk
SaaS: Software as a Service
BDaaS: Big Data as a Service
Increased Value Creation
Cloud Infrastructure
PaaS: Platform as a Service
Application Developer
IaaS: Infrastructure as a Service
System Administrator
Virtualisation Chain
1 Hardware
2 Software
3 Information & Knowledge (Big Data)

Fig. 8.3 Big data in the context of an extended service infrastructure. W. Wahlster (2013, Personal Communication)

This virtualization chain from hardware to software to information and knowledge also identifies the skills needed to maintain the infrastructure. Knowledge workers or data scientists are needed to run big data and knowledge services.

8.5.2.4 Complex Exploration

Big data exploration tools support complex datasets and their analysis through a multitude of new approaches, e.g. Sect. 8.5.1.4 on visualization. Current methods for exploration of data and analysis results have a central shortcoming in that a user can follow their exploration only selectively in one direction. If they enter a dead end or otherwise unsatisfactory state, they have to backtrack to a previous state, much as in depth-first search or hill-climbing algorithms. Emerging user interfaces for parallel exploration (CITE) are more versatile and can be compared to best-first or beam searches: the user can follow and compare multiple sequences of exploration at the same time.

Early instances of this approach have been developed under the name “subjunctive interfaces” (Lunzer and Hornbæk 2008) and applied to geographical datasets (Javed et al. 2012) and as “parallel faceted browsing” (Buschbeck et al. 2013). The latter approach assumes structured data but is applicable to all kinds of datasets, including analysis results and CEP (complex event processing).

These complex exploration tools address an inherent danger in big data analysis that arises when large datasets are automatically searched for correlations: an increasing number of seemingly statistically significant correlations will be found and need to be tested for underlying causations in a model or by expert human analysis. Complex exploration can support the checking process by allowing a parallel exploration of variations of a pattern and expected consequences of assumed causation.

8.6 Sectors Case Studies for Big Data Usage

In this section an overview of case studies that demonstrate the actual and potential value of big data usage is presented. More details can be found in Zillner et al. (2013, 2014). The use cases selected here exemplify particular aspects that are covered in those reports.

8.6.1 Healthcare: Clinical Decision Support

Description Clinical decision support (CDS) applications aim to enhance the efficiency and quality of care operations by assisting clinicians and healthcare professionals in their decision-making process. CDS applications enable contextdependent information access by providing pre-diagnosis information, or by validating and correction of data. Thus, CDS systems support clinicians in informed decision-making, which again helps to reduce treatment errors as well as helps to improve efficiency.

By relying on big data technology, future clinical decisions support applications will become substantially more intelligent. An example use case is the pre-diagnosis of medical images, with treatment recommendations reflecting existing medical guidelines.

The core prerequisite is the comprehensive data integration and the very high level of data quality necessary for physicians to actually rely on automated decision support.

8.6.2 Public Sector: Monitoring and Supervision of Online Gambling Operators

Description This future scenario represents a clear need. The main goal involved is fraud detection that is hard to execute as the amount of data received in real time, on a daily and monthly basis, cannot be processed with standard database tools. Real-time data is received from gambling operators every five minutes. Currently, supervisors have to define the cases on which to apply offline analysis of selected data.

The core prerequisite is a need to explore data interactively, compare different models and parameter settings based on technology, e.g. complex event processing that allows the real-time analysis of such a dataset. This use case relates to the issues on visual analytics and exploration, and predictive analytics.

8.6.3 Telco, Media, and Entertainment: Dynamic Bandwidth Increase

Description The introduction of new Telco offerings (e.g. a new gaming application) can cause problems with bandwidth allocations. Such scenarios are of special importance to telecommunication providers, as more profit is made with data services than with voice services. In order to pinpoint the cause of bandwidth problems, transcripts of call-centre conversations can be mined to identify customers and games involved with timing information, putting into place infrastructure measures to dynamically change the provided bandwidth according to usage.

The core prerequisites are related to predictive analysis. If problems can be detected while they are building up, peaks can be avoided altogether. Where the decision support can be automated, this scenario can be extended to prescriptive analysis.

8.6.4 *Manufacturing: Predictive Analysis*

Description Where sensor data, contextual and environmental data, is available, possible failures of machinery can be predicted. The predictions are based on abnormal sensor values that correspond to functional models of failure. Furthermore, context information such as inferences on heavy or light usage depending on the tasks executed (taken, e.g. from an ERP system) and contributing information such as weather conditions, etc., can be taken into account.

The core prerequisites, besides classical requirements such as data integration from the various, partially unstructured, data sources, are transparent prediction models and sufficiently large datasets to enable the underlying machine-learning algorithms.

8.7 Conclusions

This chapter provides state of the art as well as future requirements and emerging trends of big data usage.

The major uses of big data applications are in decision support, in predictive analytics (e.g. for predictive maintenance), and in simulation and modelling. New trends are emerging in visualization (visual analytics) and new means of exploration and comparison of alternate and competing analyses.

A special area of use cases for big data is the manufacturing, transportation, and logistics sector with a new trend "Industry 4.0". The emergence of cyber-physical systems for production, transportation, logistics, and other sectors brings new challenges for simulation and planning, for monitoring, control, and interaction (by experts and non-experts) with machinery or big data usage applications. On a larger scale, new services and a new service infrastructure are required. Under the title "smart data" and smart data services, requirements for data and also service markets are formulated. Besides the technology infrastructure for the interaction and collaboration of services from multiple sources, there are legal and regulatory issues that need to be addressed. A suitable service infrastructure is also an opportunity for SMEs to take part in big data usage scenarios by offering specific services, e.g., through data service marketplaces.

References

Anderson, C. *The end of theory*. Wired, 16.07, 2008. Available at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Apache Spark. http://spark.apache.org/, (last retrieved April 2014).

Bertolucci, J. *IBM's Predictions: 6 Big Data Trends In 2014*, December 2013. Available at http://www.informationweek.com/big-data/big-data-analytics/ibms-predictions-6-big-data-trends-in-2014-/d/d-id/1113118

Bitkom. (Ed.). (2012). *Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte*. Available at http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online%281%29.pdf

Buschbeck, S., Jameson, A., Spirescu, A., Schneeberger, T., Troncy, R., & Khrouf, H., et al. (2013). Parallel faceted browsing. In *Extended Abstracts of CHI 2013, the Conference on Human Factors in Computing Systems (Interactivity Track)*.

Chattopadhyay, B. (Google), *Youtube Data Warehouse, Latest technologies behind Youtube, including Dremel and Tenzing*, XLDB 2011, Stanford.

Chauhan, N. (2013). *Modernizing machine-to-machine interactions: A platform for Igniting the Next Industrial Revolution, GE Software*. Available at http://www.gesoftware.com/sites/default/files/GE-Software-Modernizing-Machine-to-Machine-Interactions.pdf

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.

e-skills. (2013). *Big Data Analytics: An assessment of demand for labour and skills, 2012-2017*. e-skills, London. Available at http://www.e-skills.com/research/research-publications/big-data-analytics/

Evans, P. C., Annunziata, M. (2012). *Industrial internet: Pushing the boundaries of minds and machines*, GE, November 26, 2012.

Gowing, W., & Nickson, J. (2010) Pseudonymisation Technical White Paper, NHS connecting for health, March 2010.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The 'big data' revolution in healthcare, January 2013. Available at http://www.mckinsey.com/insights/health_systems/~/media/7764A72F70184C8EA88D805092D72D58.ashx

ICO. (2012). *Anonymisation: Managing data protection risk code of practice*. Wilmslow: Information Commissioner's Office.

Javed, W., Ghani, S., & Elmqvist, N. (2012). PolyZoom: Multiscale and multifocus explora-tion in 2D visual spaces. In *Human factors in computing systems: CHI 2012 conference proceedings*. New York: ACM.

Keim, D., Kohlhammer, J., & Ellis, G. (eds.) (2010) *Mastering the information age: solving problems with visual analytics*. Eurographics Association.

Lo, S. (2012). *Big data facts and figures*, November 2012. Available at http://blogs.sap.com/innovation/big-data/big-data-facts-figures-02218

Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., & Hellerstein, J. M. (2012). Distributed Graph-Lab: A framework for machine learning and data mining in the cloud. *Proceedings of the VLDB Endowment, 5*(8), 716–727.

Lunzer, A., & Hornbæk, K. (2008). Subjunctive interfaces: Extending applications to support parallel setup, viewing and control of alternative scenarios. *ACM Transactions on Computer-Human Interaction, 14*(4), 17.

Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., et al. (2010). Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, ACM (pp. 135–146).

Manyika, J. et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company.

Markl, V., Hoeren, T., & Krcmar, H. (2013). Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen, November 2013.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of Visual Languages*.

Spence, R. (2006). *Information visualization – design for interaction* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Stratosphere Project. https://www.stratosphere.eu/, (last retrieved April 2014).

Ward, M., Grinstein, G. G., & Keim, D. (2010). *Interactive data visualization: Foundations, techniques, and applications*. Natick, MA: Taylor & Francis.

Zillner, S., Rusitschka, S., Munné, R., Lippell, H., Lobillo, F., Hussain, K., et al. (2013). *D2.3.1. First draft of the sectorial requisites*. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Zillner, S., Rusitschka, S., Munné, R., Strohbach, M., van Kasteren, T., Lippell, H., et al. (2014). *D2.3.2. Final version of the sectorial requisites*. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).