Chapter 7
Big Data Storage

Martin Strohbach, Jörg Daubert, Herman Ravkin, and Mario Lischka

7.1 Introduction

This chapter provides an overview of big data storage technologies which served as an input towards the creation of a cross-sectorial roadmap for the development of big data technologies in a range of high-impact application domains. Rather than elaborating on concrete individual technologies, this chapter provides a broad overview of data storage technologies so that the reader may get a high level understanding about the capabilities of individual technologies and areas that require further research. Consequently, the social and economic impacts are described, and selected case studies illustrating the use of big data storage technologies are provided. The full results of the analysis on big data storage can be found in Curry et al. (2014).

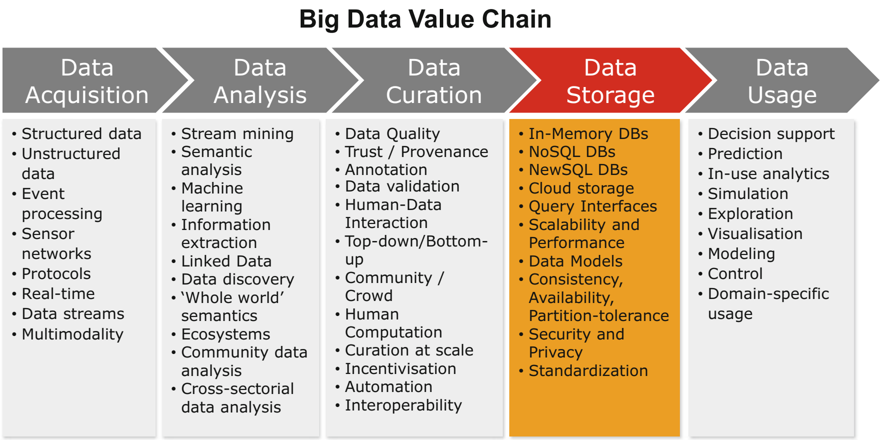
The position of big data storage within the overall big data value chain can be seen in Fig. 7.1. Big data storage is concerned with storing and managing data in a scalable way, satisfying the needs of applications that require access to the data. The ideal big data storage system would allow storage of a virtually unlimited amount of data, cope both with high rates of random write and read access, flexibly and efficiently deal with a range of different data models, support both structured and unstructured data, and for privacy reasons, only work on encrypted data.

M. Strohbach (✉) • J. Daubert
AGT International, Hilpertstr, 35, 64295 Darmstadt, Germany
e-mail: MStrohbach@agtinternational.com; jdaubert@agtinternational.com

H. Ravkin
Department of Industrial Engineering, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel
e-mail: Herman.Ravkin@gmail.com

M. Lischka
acentrix GmbH, Erika-Mann-Strasse 69, 80636 Munich, Germany
e-mail: m.lischka@acentrix.de

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_7



Big Data Value Chain

Data Acquisition
- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis
- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- 'Whole world' semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation
- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage
- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage
- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

Fig. 7.1 Data storage in the big data value chain

Obviously, all these needs cannot be fully satisfied. But over recent years many new storage systems have emerged that at least partly address these challenges.[1]

This chapter provides an overview of big data storage technologies and identifies some areas where further research is required. *Big data storage technologies* are referred to as storage technologies that in some way specifically address the volume, velocity, or variety challenge and do *not* fall in the category of relational database systems. This does not mean that relational database systems do not address these challenges, but alternative storage technologies such as columnar stores and clever combinations of different storage systems, e.g. using the Hadoop Distributed File System (HDFS), are often more efficient and less expensive (Marz and Warren 2014).

Big data storage systems typically address the volume challenge by making use of distributed, shared nothing architectures. This allows addressing increased storage requirements by scaling out to new nodes providing computational power and storage. New machines can seamlessly be added to a storage cluster and the storage system takes care of distributing the data between individual nodes transparently.

Storage solutions also need to cope with the velocity and variety of data. Velocity is important in the sense of query latencies, i.e. how long does it take to get a reply for a query? This is particularly important in the face of high rates of incoming data. For instance, random write access to a database can slow down query performance considerably if it needs to provide transactional guarantees. In contrast, variety relates to the level of effort that is required to integrate and work with data that originates from a large number of different sources. For instance, graph databases are suitable storage systems to address these challenges.

[1] See for instance the map of 451 Research available at https://451research.com/state-of-the-database-landscape

Section 7.2 summarizes key insights and Sect. 7.3 illustrates the social and economic impact of data storage. Section 7.4 presents the current state-of-the-art including storage technologies and solutions for security and privacy. Section 7.5 includes future requirements and emerging trends for data storage that will play an important role for unlocking the value hidden in large datasets. Section 7.6 presents three selected case studies, and the chapter is concluded in Sect. 7.7.

7.2 Key Insights for Big Data Storage

As a result of the analysis of current and future data storage technologies, a number of insights were gained relating to data storage technologies. It became apparent that big data storage has become a commodity business and that scalable storage technologies have reached an enterprise-grade level that can manage virtually unbounded volumes of data. Evidence is provided by the widespread use of Hadoop-based solutions offered by vendors such as Cloudera (2014a), Hortonworks (2014), and MapR (2014) as well as various NoSQL[2] database vendors, in particular those that use in-memory and columnar storage technologies. Compared to traditional relational database management systems that rely on row-based storage and expensive caching strategies, these novel big data storage technologies offer better scalability at lower operational complexity and costs.

Despite these advances that improve the performance, scalability, and usability of storage technologies, there is still significant untapped potential for big data storage technologies, both for using and further developing the technologies:

- **Potential to Transform Society and Businesses across Sectors:** Big data storage technologies are a key enabler for advanced analytics that have the potential to transform society and the way key business decisions are made. This is of particular importance in traditionally non-IT-based sectors such as energy. While these sectors face non-technical issues such as the lack of skilled big data experts and regulatory barriers, novel data storage technologies have the potential to enable new value-generating analytics in and across various industrial sectors.
- **Lack of Standards Is a Major Barrier:** The history of NoSQL is based on solving specific technological challenges which lead to a range of different storage technologies. The large range of choices coupled with the lack of standards for querying the data makes it harder to exchange data stores as it may tie application specific code to a certain storage solution.
- **Open Scalability Challenges in Graph-Based Data Stores:** Processing data based on graph data structures is beneficial in an increasing amount of applications. It allows better capture of semantics and complex relationships with other

[2] NoSQL is typically referred to as “Not only SQL”.

pieces of information coming from a large variety of different data sources, and has the potential to improve the overall value that can be generated by analysing the data. While graph databases are increasingly used for this purpose, it remains hard to efficiently distribute graph-based data structure across computing nodes.

- **Privacy and Security Is Lagging Behind:** Although there are several projects and solutions that address privacy and security, the protection of individuals and securing their data lags behind the technological advances of data storage systems. Considerable research is required to better understand how data can be misused, how it needs to be protected and integrated in big data storage solutions.

7.3 Social and Economic Impact of Big Data Storage

As emerging big data technologies and their use in different sectors show, the capability to store, manage, and analyse large amounts of heterogeneous data hints towards the emergence of a data-driven society and economy with huge transformational potential (Manyika et al. 2011). Enterprises can now store and analyse more data at a lower cost while at the same time enhancing their analytical capabilities. While companies such as Google, Twitter, and Facebook are established players for which data constitutes the key asset, other sectors also tend to become more data driven. For instance, the health sector is an excellent example that illustrates how society can expect better health services by better integration and analysis of health-related data (iQuartic 2014).

Many other sectors are heavily impacted by the maturity and cost-effectiveness of technologies that are able to handle big datasets. For instance, in the media sector the analysis of social media has the potential to transform journalism by summarizing news created by a large amount of individuals. In the transport sector, the consolidated data management integration of transport systems has the potential to enable personalized multimodal transportation, increasing the experience of travellers within a city and at the same time helping decision-makers to better manage urban traffic. In all of these areas, NoSQL storage technologies prove a key enabler to efficiently analyse large amounts of data and create additional business value.

On a cross-sectorial level, the move towards a data-driven economy can be seen by the emergence of data platforms such as datamarket.com (Gislason 2013), infochimp.com, and open data initiatives of the European Union such as open-data.europa.eu and other national portals (e.g. data.gov, data.gov.uk, data.gov.sg) (Ahmadi Zeleti et al. 2014). Technology vendors are supporting the move towards a data-driven economy as can be seen by the positioning of their products and services. For instance, Cloudera is offering a product called the *enterprise data hub* (Cloudera 2014b), an extended Hadoop ecosystem that is positioned as a data management and analysis integration point for the whole company.

Further to the benefits described above, there are also threats to big data storage technologies that must be addressed to avoid any negative impact. This relates for instance to the challenge of protecting the data of individuals and reducing the energy consumption of data centres (Koomey 2008).

7.4 Big Data Storage State-of-the-Art

This section provides an overview of the current state-of-the-art in big data storage technologies. Section 7.4.1 describes the storage technologies, and Sect. 7.4.2 presents technologies related to secure and privacy-preserving data storage.

7.4.1 Data Storage Technologies

During the last decade, the need to deal with the data explosion (Turner et al. 2014) and the hardware shift from *scale-up* to *scale-out* approaches led to an explosion of new big data storage systems that shifted away from traditional relational database models. These approaches typically sacrifice properties such as data consistency in order to maintain fast query responses with increasing amounts of data. Big data stores are used in similar ways as traditional relational database management systems, e.g. for online transactional processing (OLTP) solutions and data warehouses over structured or semi-structured data. Particular strengths are in handling unstructured and semi-structured data at large scale.

This section assesses the current state-of-the-art in data store technologies that are capable of handling large amounts of data, and identifies data store related trends. Following are differing types of storage systems:

- **Distributed File Systems:** File systems such as the Hadoop File System (HDFS) (Shvachko et al. 2010) offer the capability to store large amounts of unstructured data in a reliable way on commodity hardware. Although there are file systems with better performance, HDFS is an integral part of the Hadoop framework (White 2012) and has already reached the level of a de-facto standard. It has been designed for large data files and is well suited for quickly ingesting data and bulk processing.
- **NoSQL Databases:** Probably the most important family of big data storage technologies are NoSQL database management systems. NoSQL databases use data models from outside the relational world that do not necessarily adhere to the transactional properties of atomicity, consistency, isolation, and durability (ACID).
- **NewSQL Databases:** A modern form of relational databases that aim for comparable scalability as NoSQL databases while maintaining the transactional guarantees made by traditional database systems.
- **Big Data Querying Platforms:** Technologies that provide query facades in front of big data stores such as distributed file systems or NoSQL databases. The

main concern is providing a high-level interface, e.g. via SQL[3] like query languages and achieving low query latencies.

7.4.1.1 NoSQL Databases

NoSQL databases are designed for scalability, often by sacrificing consistency. Compared to relational databases, they often use low-level, non-standardized query interfaces, which make them more difficult to integrate in existing applications that expect an SQL interface. The lack of standard interfaces makes it harder to switch vendors. NoSQL databases can be distinguished by the data models they use.

- **Key-Value Stores:** Key-value stores allow storage of data in a schema-less way. Data objects can be completely unstructured or structured and are accessed by a single key. As no schema is used, it is not even necessary that data objects share the same structure.
- **Columnar Stores:** According to Wikipedia "A column-oriented DBMS is a database management system (DBMS) that stores data tables as sections of columns of data rather than as rows of data, like most relational DBMSs" (Wikipedia 2013). Such databases are typically sparse, distributed, and persistent multi-dimensional sorted maps in which data is indexed by a triple of a row key, column key, and a timestamp. The value is represented as an uninterrupted string data type. Data is accessed by column families, i.e. a set of related column keys that effectively compress the sparse data in the columns. Column families are created before data can be stored and their number is expected to be small. In contrast, the number of columns is unlimited. In principle columnar stores are less suitable when all columns need to be accessed. However in practice this is rarely the case, leading to superior performance of columnar stores.
- **Document Databases:** In contrast to the values in a key-value store, documents are structured. However, there is no requirement for a common schema that all documents must adhere to as in the case for records in relational databases. Thus document databases are referred to as storing semi-structured data. Similar to key-value stores, documents can be queried using a unique key. However, it is possible to access documents by querying their internal structure, such as requesting all documents that contain a field with a specified value. The capability of the query interface is typically dependent on the encoding format used by the databases. Common encodings include XML or JSON.
- **Graph Databases:** Graph databases, such as Neo4J (2015), store data in graph structures making them suitable for storing highly associative data such as social network graphs. A particular flavour of graph databases are triple stores such as AllegroGraph (Franz 2015) and Virtuoso (Erling 2009) that are specifically

[3] Here and throughout this chapter SQL refers to the Standard Query Language as defined in the ISO/IEC Standard 9075-1:2011.

designed to store RDF triples. However, existing triple store technologies are not yet suitable for storing truly large datasets efficiently.

While in general NoSQL data stores scale better than relational databases, scalability decreases with increased complexity of the data model used by the data store. This particularly applies to graph databases that support applications that are both write and read intensive. One approach to optimize read access is to partition the graph into sub-graphs that are minimally connected between each other and to distribute these sub-graphs between computational nodes. However, as new edges are added to a graph the connectivity between sub-graphs may increase considerably. This may lead to higher query latencies due to increased networks traffic and non-local computations. Efficient *sharding schemes* must therefore carefully consider the overhead required for dynamically re-distributing graph data.

7.4.1.2 NewSQL Databases

NewSQL databases are a modern form of relational databases that aim for comparable scalability with NoSQL databases while maintaining the transactional guarantees made by traditional database systems. According to Venkatesh and Nirmala (2012) they have the following characteristics:

- SQL is the primary mechanism for application interaction
- ACID support for transactions
- A non-locking concurrency control mechanism
- An architecture providing much higher per-node performance
- A scale-out, shared-nothing architecture, capable of running on a large number of nodes without suffering bottlenecks

The expectation is that NewSQL systems are about 50 times faster than traditional OLTP RDBMS. For example, VoltDB (2014) scales linearly in the case of non-complex (single-partition) queries and provides ACID support. It scales for dozens of nodes where each node is restricted to the size of the main memory.

7.4.1.3 Big Data Query Platforms

Big data query platforms provide query facades on top of underlying big data stores that simplify querying the underlying data stores. They typically offer an SQL-like query interface for accessing the data, but differ in their approach and performance.

Hive (Thusoo et al. 2009) provides an abstraction on top of the Hadoop Distributed File System (HDFS) that allows structured files to be queried by an SQL-like query language. Hive executes the queries by translating queries in MapReduce jobs. As a consequence, Hive queries have a high latency even for small datasets. Benefits of Hive include the SQL-like query interface and the flexibility to evolve schemas easily. This is possible as the schema is stored independently from the data and the data is only validated at query time. This approach is referred to as schema-on-read compared to the schema-on-write approach of SQL databases. Changing the schema is therefore a comparatively cheap operation. The Hadoop columnar store HBase is also supported by Hive.

In contrast to Hive, Impala (Russel 2013) is designed for executing queries with low latencies. It re-uses the same metadata and SQL-like user interface as Hive but uses its own distributed query engine that can achieve lower latencies. It also supports HDFS and HBase as underlying data stores.

Spark SQL (Shenker et al. 2013) is another low latency query façade that supports the Hive interface. The project claims that "it can execute Hive QL queries up to 100 times faster than Hive without any modification to the existing data or queries" (Shenker et al. 2013). This is achieved by executing the queries using the Spark framework (Zaharia et al. 2010) rather than Hadoop's MapReduce framework.

Finally, Drill is an open source implementation of Google's Dremel (Melnik et al. 2002) that similar to Impala is designed as a scalable, interactive ad-hoc query system for nested data. Drill provides its own SQL-like query language DrQL that is compatible with Dremel, but is designed to support other query languages such as the Mongo Query Language. In contrast to Hive and Impala, it supports a range of schema-less data sources, such as HDFS, HBase, Cassandra, MongoDB, and SQL databases.

7.4.1.4 Cloud Storage

As cloud computing grows in popularity, its influence on big data grows as well. While Amazon, Microsoft, and Google build on their own cloud platforms, other companies including IBM, HP, Dell, Cisco, Rackspace, etc., build their proposal around OpenStack, an open source platform for building cloud systems (OpenStack 2014).

According to IDC (Grady 2013), by 2020 40 % of the digital universe "will be 'touched' by cloud computing", and "perhaps as much as 15 % will be maintained in a cloud".

Cloud in general, and particularly cloud storage, can be used by both enterprises and end users. For end users, storing their data in the cloud enables access from everywhere and from every device in a reliable way. In addition, end users can use cloud storage as a simple solution for online backup of their desktop data. Similarly for enterprises, cloud storage provides flexible access from multiple locations and quick and easy scale capacity (Grady 2013) as well as cheaper storage prices and better support based on economies of scale (CloudDrive 2013) with cost effectiveness especially high in an environment where enterprise storage needs are changing over time up and down.

Technically cloud storage solutions can be distinguished between object and block storage. Object storage "is a generic term that describes an approach to addressing and manipulating discrete units of storage called objects" (Margaret

Rouse 2014a). In contrast, block storage data is stored in volumes also referred to as blocks. According to Margaret Rouse (2014b), “each block acts as an individual hard drive” and enables random access to bits and pieces of data thus working well with applications such as databases.

In addition to object and block storage, major platforms provide support for relational and non-relational database-based storage as well as in-memory storage and queue storage. In cloud storage, there are significant differences that need to be taken into account in the application-planning phase:

- As cloud storage is a service, applications using this storage have less control and may experience decreased performance as a result of networking. These performance differences need to be taken into account during design and implementation stages.
- Security is one of the main concerns related to public clouds. As a result the Amazon CTO predicts that in five years all data in the cloud will be encrypted by default (Vogels 2013).
- Feature rich clouds like AWS supports calibration of latency, redundancy, and throughput levels for data access, thus allowing users to find the right trade-off between cost and quality.

Another important issue when considering cloud storage is the supported consistency model (and associated scalability, availability, partition tolerance, and latency). While Amazon’s Simple Storage Service (S3) supports eventual consistency, Microsoft Azure blob storage supports strong consistency and at the same time high availability and partition tolerance. Microsoft uses two layers: (1) a stream layer “which provides high availability in the face of network partitioning and other failures”, and (2) a partition layer which “provides strong consistency guarantees” (Calder et al. 2011).

7.4.2 Privacy and Security

Privacy and security are well-recognized challenges in big data. The CSA Big Data Working Group published a list of *Top 10 Big Data Security and Privacy Challenges* (Mora et al. 2012). The following are five of those challenges that are vitally important for big data storage.

7.4.2.1 Security Best Practices for Non-relational Data Stores

The security threats for NoSQL databases are similar to traditional RDBMS and therefore the same best practices should be applied (Winder 2012). However, many security measures that are implemented by default within traditional RDBMS are missing in NoSQL databases (Okman et al. 2011). Such measures would include encryption of sensitive data, sandboxing of processes, input validation, and strong user authentication.

Some NoSQL suppliers recommend the use of databases in a trusted environment with no additional security or authentication measures in place. However, this approach is hardly reasonable when moving big data storage to the cloud.

Security of NoSQL databases is getting more attention by security researchers and hackers, and security will further improve as the market matures. For example, there are initiatives to provide access control capabilities for NoSQL databases based on Kerberos authentication modules (Winder 2012).

7.4.2.2 Secure Data Storage and Transaction Logs

Particular security challenges for data storage arise due to the distribution of data. With auto-tiering, operators give away control of data storage to algorithms in order to reduce costs. Data whereabouts, tier movements, and changes have to be accounted for by transaction log.

Auto-tiering strategies have to be carefully designed to prevent sensitive data being moved to less secure and thus cheaper tiers; monitoring and logging mechanisms should be in place in order to have a clear view on data storage and data movement in auto-tiering solutions (Mora et al. 2012).

Proxy re-encryption schemes (Blaze et al. 2006) can be applied to multi-tier storage and data sharing in order to ensure seamless confidentiality and authenticity (Shucheng et al. 2010). However, performance has to be improved for big data applications. Transaction logs for multi-tier operations systems are still missing.

7.4.2.3 Cryptographically Enforced Access Control and Secure Communication

Today, data is often stored unencrypted, and access control solely depends on a gate-like enforcement. However, data should only be accessible by authorized entities by the guarantees of cryptography—likewise in storage as well as in transmission. For these purposes, new cryptographic mechanisms are required that provide the required functionalities in an efficient and scalable way.

While cloud storage providers are starting to offer encryption, cryptographic key material should be generated and stored at the client and never handed over to the cloud provider. Some products add this functionality to the application layer of big data storage, e.g., *zNcrypt*, *Protegrity Big Data Protection for Hadoop*, and the *Intel Distribution for Apache Hadoop* (now part of *Cloudera*).

Attribute-based encryption (Goyal et al. 2006) is a promising technology to integrate cryptography with access control for big data storage (Kamara and Lauter 2010; Lee et al. 2013; Li et al. 2013).

7.4.2.4 Security and Privacy Challenges for Granular Access Control

Diversity of data is a major challenge due to equally diverse security requirements, e.g., legal restrictions, privacy policies, and other corporate policies. Fine-grained access control mechanisms are needed to assure compliance with these requirements.

Major big data components use Kerberos (Miller et al. 1987) in conjunction with token-based authentication, and Access Control Lists (ACL) based upon users and jobs. However, more fine-grained mechanism, for instance Attribute-Based Access Control (ABAC) and eXtensible Access Control Markup Language (XACLM), are required to model the vast diversity of data origins and analytical usages.

7.4.2.5 Data Provenance

Integrity and history of data objects within value chains is crucial. Traditional provenance governs mostly ownership and usage. With big data however, the complexity of provenance metadata will increase (Glavic 2014).

Initial efforts have been made to integrate provenance into the big data ecosystem (Ikeda et al. 2011; Sherif et al. 2013); however, secure provenance requires guarantees of integrity and confidentiality of provenance data in all forms of big data storage and remains an open challenge. Furthermore, the analysis of very large provenance graphs is computationally intensive and requires fast algorithms.

7.4.2.6 Privacy Challenges in Big Data Storage

Researchers have shown (Acquisti and Gross 2009) that big data analysis of publicly available information can be exploited to guess the social security number of a person. Some products selectively encrypt data fields to create reversible anonymity, depending on the access privileges.

Anonymizing and de-identifying data may be insufficient as the huge amount of data may allow for re-identification. A roundtable discussion (Bollier and Firestone 2010) advocated *transparency on the handling of data and algorithms* as well as a *new deal on big data* (Wu and Guo 2013) to empower the end user as the owner of the data. Both options not only involve organization transparency, but also technical tooling such as *Security & Privacy by Design* and the results of the EEXCESS EU FP7 project (Hasan et al. 2013).

7.5 Future Requirements and Emerging Paradigms for Big Data Storage

This section provides an overview of future requirements and emerging trends.

7.5.1 Future Requirements for Big Data Storage

Three key areas have been identified that can be expected to govern future big data storage technologies. This includes standardization of query interfaces, increasing support for data security, protection of users' privacy, and the support of semantic data models.

7.5.1.1 Standardized Query Interfaces

In the medium to long-term NoSQL databases would greatly benefit from standardized query interfaces, similar to SQL for relational systems. Currently no standards exist for the individual NoSQL storage types beyond de-facto standard APIs for graph databases (Blueprints 2014) and the SPARQL data manipulation language (Aranda et al. 2013) supported by triplestore's vendors. Other NoSQL databases usually provide their own declarative language or API, and standardization for these declarative languages is missing.

While for some database categories (key/value, document, etc.) declarative language standardization is still missing, there are efforts discussing standardization needs. For instance the ISO/IEC JTC Study Group on big data has recently recommended that existing ISO/IEC standards committee should further investigate the "definition of standard interfaces to support non-relational data stores" (Lee et al. 2014).

The definition of standardized interfaces would enable the creation of a data virtualization layer that would provide an abstraction of heterogeneous data storage systems as they are commonly used in big data use cases. Some requirements of a data virtualization layer have been discussed online in an Infoworld blog article (Kobielus 2013).

7.5.1.2 Security and Privacy

Interviews were conducted with consultants and end users of big data storage who have responsibility for security and privacy, to gain their personal views and insights. Based upon these interviews and the gaps identified in Sect. 7.4.2, several future requirements for security and privacy in big data storage were identified.

Data Commons and Social Norms Data stored in large quantities will be subject to sharing as well as derivative work in order to maximize big data benefits. Today, users are not aware how big data processes their data (transparency), and it is not clear how big data users can share and obtain data efficiently. Furthermore, legal constraints with respect to privacy and copyright in big data are currently not completely clear within the EU. For instance, big data allows novel analytics based upon aggregated data from manifold sources. How does this approach affect private information? How can rules and regulations for remixing and derivative works be applied to big data? Such uncertainty may lead to a disadvantage of the EU compared to the USA.

Data Privacy Big data storage must comply with EU privacy regulations such as Directive 95/46/EC when personal information is being stored. Today, heterogeneous implementations of this directive render the storage of personal information in big data difficult. The General Data Protection Regulation (GDRP)—first proposed in 2012—is an on-going effort to harmonize data protection among EU member states. The GDRP is expected to influence future requirements for big data storage. As of 2014, the GDRP is subject to negotiations that make it difficult to estimate the final rules and start of enforcement. For instance, the 2013 draft version allows *data subjects* (persons) to request data controllers to delete personal data, which is often not sufficiently considered by big data storage solutions.

Data Tracing and Provenance Tracing and provenance of data is becoming more and more important in big data storage for two reasons: (1) users want to understand where data comes from, if the data is correct and trustworthy, and what happens to their results and (2) big data storage will become subject to compliance rules as big data enters critical business processes and value chains. Therefore, big data storage has to maintain provenance metadata, provide provenance along the data processing chain, and offer user-friendly ways to understand and trace the usage of data.

Sandboxing and Virtualization Sandboxing and virtualization of big data analytics becomes more important in addition to access control. According to economies of scale, big data analytics benefit from resource sharing. However, security breaches of shared analytical components lead to compromised cryptographic access keys and full storage access. Thus, jobs in big data analytics must be sandboxed to prevent an escalation of security breaches and therefore unauthorized access to storage.

7.5.1.3 Semantic Data Models

The multitude of heterogeneous data sources increases development costs, as applications require knowledge about individual data formats of each individual source. An emerging trend is the semantic web and in particular the semantic sensor web that tries to address this challenge. A multitude of research projects are concerned with all levels of semantic modelling and computation. As detailed in this book, the need for semantic annotations has for instance been identified for the health sector. The requirement for data storage is therefore to support the large-scale storage and management of semantic data models. In particular trade-offs between expressivity and efficient storage and querying need to be further explored.

7.5.2 Emerging Paradigms for Big Data Storage

There are several new paradigms emerging for the storage of large and complex datasets. These new paradigms include, among others, the increased use of NoSQL databases, convergence with analytics frameworks, and managing data in a central data hub.

7.5.2.1 Increased Use of NoSQL Databases

NoSQL databases, most notably graph databases and columnar stores, are increasingly used as a replacement or complement to existing relational systems.

For instance, the requirement of using semantic data models and cross linking data with many different data and information sources strongly drives the need to be able to store and analyse large amounts of data using graph-based models. However, this requires overcoming the limitation of current graph-based systems as described above. For instance, Jim Webber states "Graph technologies are going to be incredibly important" (Webber 2013). In another interview, Ricardo Baeza-Yates, VP of Research for Europe and Latin America at Yahoo!, also states the importance of handling large-scale graph data (Baeza-Yates 2013). The Microsoft research project Trinity achieved a significant breakthrough in this area. Trinity is an in-memory data storage and distributed processing platform. By building on its very fast graph traversal capabilities, Microsoft researchers introduced a new approach to cope with graph queries. Other projects include Google's knowledge graph and Facebook's graph search that demonstrate the increasing relevance and growing maturity of graph technologies.

7.5.2.2 In-Memory and Column-Oriented Designs

Many modern high-performance NoSQL databases are based on columnar designs. The main advantage is that in most practical applications only a few columns are needed to access the data. Consequently storing data in columns allows faster access. In addition, column-oriented databases often do not support the expensive join operations from the relational world. Instead, a common approach is to use a single wide column table that stores the data based on a fully denormalized schema.

According to Michael Stonebraker "SQL vendors will all move to column stores, because they are wildly faster than row stores" (Stonebraker 2012a).

High-performance in-memory databases such as SAP HANA typically combine in-memory techniques with column-based designs. In contrast to relational systems that cache data in-memory, in-memory databases can use techniques such as anti-caching (DeBrabant et al. 2013). Harizopoulos et al. have shown that the most time for executing a query is spent on administrative tasks such as buffer management and locking (Harizopoulos et al. 2008).

7.5.2.3 Convergence with Analytics Frameworks

During the course of the project many scenarios have been identified that call for better analysis of available data to improve operations in various sectors. Technically, this means an increased need for complex analytics that goes beyond simple aggregations and statistics. Stonebraker points out that the need for complex analytics will strongly impact existing data storage solutions (Stonebraker 2012b).

As use case specific analytics are one of the most crucial components that are creating actual business value, it becomes increasingly important to scale up these analytics satisfying performance requirements, but also to reduce the overall development complexity and cost. Figure 7.2 shows some differences between using separate systems for data management and analytics versus integrated analytical databases.

7.5.2.4 The Data Hub

A central data hub that integrates all data in an enterprise is a paradigm that considers managing all company data as a whole, rather than in different, isolated databases managed by different organizational units. The benefit of a central data hub is that data can be analysed as a whole, linking various datasets owned by the company thus leading to deeper insights.

Typical technical implementations are based on a Hadoop-based system that may use HDFS or HBase (Apache 2014) to store an integrated master dataset. On one hand, this master dataset can be used as ground truth and backup for existing data management systems, but it also provides the basis for advanced analytics that combine previously isolated datasets.

Companies such as Cloudera use this paradigm to market their Hadoop distribution (Cloudera 2014b). Many use cases of enterprise data hub exist already. A case study in the financial sector is described in the next section.

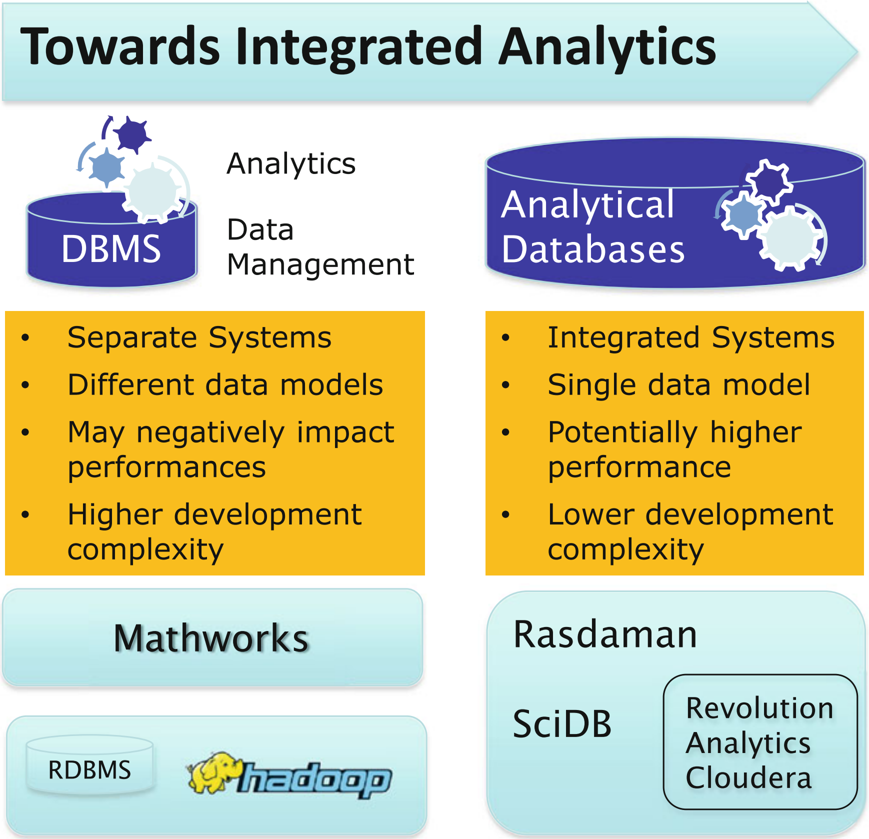


Fig. 7.2 Paradigm shift from pure data storage systems to integrated analytical databases

7.6 Sector Case Studies for Big Data Storage

In this section three selected use cases are described that illustrate the potential and need for future storage technologies. The health use case illustrates how social media based analytics is enabled by NoSQL storage technologies. The second use case from the financial sector illustrates the emerging paradigm of a centralized data hub. The last use case from the energy sector illustrates the benefits of managing fine-grained Internet of Things (IoT) data for advanced analytics. An overview of the key characteristics of the use case can be found in Table 7.1. More case studies are presented in Curry et al. (2014).

Table 7.1 Key characteristics of selected big data storage case studies

Case study	Sector	Volume	Storage technologies	Key requirements
Treato: Social media based medication intelligence	Health	>150 TB	HBase	Cost-efficiency, scalability limitations of relational DBs
Centralized data hub	Finance	Between several petabytes and over 150 PB	Hadoop/ HDFS	Building more accurate models, scale of data, suitability for unstructured data
Smart grid	Energy	Tens of TB per day	Hadoop	Data volume, operational challenges

7.6.1 Health Sector: Social Media-Based Medication Intelligence

Treato is an Israeli company that specializes in mining user-generated content from blogs and forums in order to provide brand intelligence services to pharmaceutical companies. As Treato is analysing the social web, it falls into the "classical" category of analysing large amounts of unstructured data, an application area that often asks for big data storage solutions. Treato's service as a use case demonstrates the value of using big data storage technologies. The information is based on a case study published by Cloudera (2012), the company that provided the Hadoop distribution Treato has been using.

While building its prototype, Treato discovered "that side effects could be identified through social media long before pharmaceutical companies or the Food & Drug Administration (FDA) issued warnings about them. For example, when looking at discussions about Singulair, an asthma medication, Treato found that almost half of UGC discussed mental disorders; the side effect would have been identifiable four years before the official warning came out." (Cloudera 2012).

Treato initially faced two major challenges: First, it needed to develop the analytical capabilities to analyse patient's colloquial language and map that into a medical terminology suitable for delivering insights to its customers. Second, it was necessary to analyse large amounts of data sources as fast as possible in order to provide accurate information in real time.

The first challenge, developing the analytics, has been addressed initially with a non-Hadoop system based on a relational database. With that system Treato was facing the limitation that it could only handle "data collection from dozens of websites and could only process a couple of million posts per day" (Cloudera 2012). Thus, Treato was looking for a cost-efficient analytics platform that could fulfil the following key requirements:

1. Reliable and scalable storage
2. Reliable and scalable processing infrastructure
3. Search engine capabilities for retrieving posts with high availability
4. Scalable real-time store for retrieving statistics with high availability

As a result Treato decided on a Hadoop-based system that uses HBase to store the list of URLs to be fetched. The posts available at these URLs are analysed by using natural language processing in conjunction with their proprietary ontology. In addition "each individual post is indexed, statistics are calculated, and HBase tables are updated" (Cloudera 2012).

According to the case study report, the Hadoop-based solution stores more than 150 TB of data including 1.1 billion online posts from thousands of websites including about more than 11,000 medications and more than 13,000 conditions. Treato is able to process 150–200 million user posts per day.

For Treato, the impact of the Hadoop-based storage and processing infrastructure is that they obtain a scalable, reliable, and cost-effective system that may even create insights that would not have been possible without this infrastructure. The case study claims that with Hadoop, Treato improved execution time at least by a factor of six. This allowed Treato to respond to a customer request about a new medication within one day.

7.6.2 Finance Sector: Centralized Data Hub

As mapped out in the description of the sectorial roadmaps (Lobillo et al. 2013), the financial sector is facing challenges with respect to increasing data volumes and a variety of new data sources such as social media. Here use cases are described for the financial sector based on a Cloudera solution brief (Cloudera 2013).

Financial products are increasingly digitalized including online banking and trading. As online and mobile access simplifies access to financial products, there is an increased level of activity leading to even more data. The potential of big data in this scenario is to use all available data for building accurate models that can help the financial sector to better manage financial risks. According to the solution brief, companies have access to several petabytes of data. According to Larry Feinsmith, managing director of JPMorgan Chase, his company is storing over 150 petabytes online and use Hadoop for fraud detection (Cloudera 2013).

Secondly, new data sources add to both the volume and variety of available data. In particular, unstructured data from weblogs, social media, blogs, and other news feeds can help in customer relationship management, risk management, and maybe even algorithmic trading (Lobillo et al. 2013). Pulling all the data together in a centralized data hub enables more detailed analytics that can provide a competitive edge. However traditional systems cannot keep up with the scale, costs, and cumbersome integration of traditional extract, transform, load (ETL) processes using fixed data schemes, nor are they able to handle unstructured data. Big data storage systems however scale extremely well and can process both structured and unstructured data.

7.6.3 Energy: Device Level Metering

In the energy sector, smart grid and smart meter management is an area that promises both high economic and environmental benefits. As depicted in Fig. 7.3, the introduction of renewable energies such as photovoltaic systems deployed on houses can cause grid instabilities. Currently grid operators have little knowledge about the last mile to energy consumers. Thus they are not able to appropriately react to instabilities caused at the very edges of the grid network. By analysing smart meter data sampled at second intervals, short-term forecasting of energy demands and managing the demand of devices such as heating and electrical cars becomes possible, thus stabilizing the grid. If deployed in millions of households the data volumes can reach petabyte scale, thus greatly benefiting from new storage technologies. Table 7.2 shows the data volume only for the raw data collected for one day.

The Peer Energy Cloud (PEC) project (2014) is a public funded project that has demonstrated how smart meter data can be analysed and used for trading energy in the local neighbourhood, thus increasing the overall stability of the power grid. Moreover, it has successfully shown that by collecting more fine granular data, i.e. monitoring energy consumption of individual devices in the household, the accuracy of predicting the energy consumption of households can be significantly improved (Ziekow et al. 2013). As the data volumes increase it becomes

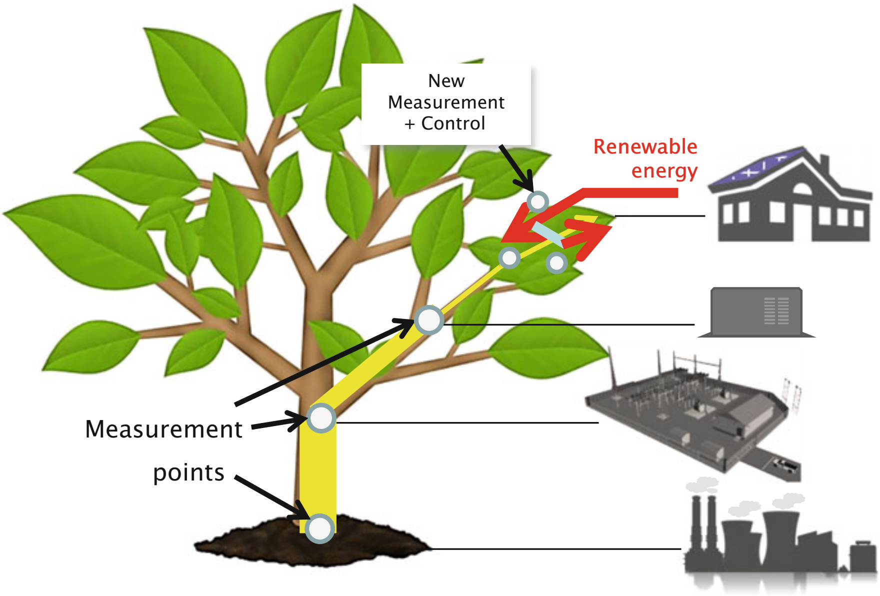


Fig. 7.3 Introduction of renewable energy at consumer sites changes the topology of the energy grid and requires new measurement points at the leaves of the grid

Table 7.2 Calculation of the amount of data sampled by smart meters

Sampling rate	1 Hz
Record size	50 Bytes
Raw data per day and household	4.1 MB
Raw data per day for 10 Mio customers	~39 TB

increasingly difficult to handle the data with legacy relational databases (Strohbach et al. 2011).

7.7 Conclusions

The chapter contains an overview of current big data storage technologies as well as emerging paradigms and future requirements. The overview specifically included technologies and approaches related to privacy and security. Rather than focusing on detailed descriptions of individual technologies a broad overview was provided, and technical aspects that have an impact on creating value from large amounts of data highlighted. The social and economic impact of big data storage technologies was described, and three selected case studies in three different sectors were detailed, which illustrate the need for easy to use scalable technologies.

It can be concluded that there is already a huge offering of big data storage technologies. They have reached a maturity level that is high enough that early adopters in various sectors already use or plan to use them. Big data storage often has the advantage of better scalability at a lower price tag and operational complexity. The current state of the art reflects that the efficient management of almost any size of data is not a challenge per se. Thus it has huge potential to transform business and society in many areas.

It can also be concluded that there is a strong need to increase the maturity of storage technologies so that they fulfil future requirements and lead to a wider adoption, in particular in non-IT-based companies. The required technical improvements include the scalability of graph databases that will enable better handling of complex relationships, as well as further minimizing query latencies to big datasets, e.g. by using in-memory databases. Another major roadblock is the lack of standardized interfaces to NoSQL database systems. The lack of standardization reduces flexibility and slows down adoption. Finally, considerable improvements for security and privacy are required. Secure storage technologies need to be further developed to protect the privacy of users.

More details about big data storage technologies can be found in Curry et al. (2014). This report, in conjunction with the analysis of the public and 10 industrial sectors (Zillner et al. 2014), has been used as a basis to develop the cross-sectorial roadmap described in this book.

References

Acquisti, A., & Gross, R. (2009). Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences, 106*(27), 10975–10980.

Ahmadi Zeleti, F., Ojo, A., & Curry, E. (2014). Business models for the open data industry: Characterization and analysis of emerging models. In *15th Annual International Conference on Digital Government Research* (dg.o 2014) (pp. 215–226). ACM.

Apache. (2014). Apache HBase Project Website. http://hbase.apache.org. Accessed Nov 21, 2014.

Aranda, C. B., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., et al. (2013). SPARQL 1.1 overview. In *W3C Recommendation*.

Baeza-Yates, R. (2013). BIG Interview by John Dominque.

Blaze, M., Bleumer, G., & Strauss, M. (2006). Divertible protocols and atomic proxy cryptography. In *Proceedings of Eurocrypt* (pp. 127–144).

Blueprints. (2014). Blueprints Project Homepage. https://github.com/tinkerpop/blueprints/wiki. Accessed Feb 4, 2015.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data*. Washington, DC: Aspen Institute, Communications and Society Program.

Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolvsvold, A., McKelvie, S. et al. (2011). Windows azure storage: a highly available cloud storage service with strong consistency. In: *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles* (143–157). New York, NY: ACM.

CloudDrive. (2013). Advantages of cloud data storage. http://www.clouddrive.com.au/download/www.clouddrive.com.au-WhitePaper.pdf. Accessed Nov 20, 2013.

Cloudera. (2012). Treato Customer Case Study. https://www.cloudera.com/content/dam/cloudera/Resources/PDF/casestudy/Cloudera_Customer_Treato_Case_Study.pdf

Cloudera. (2013). *Identifying fraud, managing risk and improving compliance in financial services*.

Cloudera. (2014a). Cloudera Company Web page. www.cloudera.com. Accessed May 6, 2015.

Cloudera. (2014b). Rethink data. http://www.cloudera.com/content/cloudera/en/new/. Accessed Feb 4, 2014.

Curry, E., Ngonga, A., Domingue, J., Freitas, A., Strohbach, M., Becker, T., et al. (2014). D2.2.2. Final version of the technical white paper. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

DeBrabant, J., Pavlo, A., Tu, S., Stonebraker, M., & Zdonik, S. (2013). Anti-caching: a new approach to database management system architecture. In *Proceedings of the VLDB Endowment* (pp 1942–1953).

Erling, O. (2009). Virtuoso, a Hybrid RDBMS/Graph Column Store. In: Virtuoso Open-Source Wiki. http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSArticleVirtuosoAHybridRDBMSGraphColumnStore. Accessed Feb 6, 2015.

Franz. (2015). Allegrograph product web page. http://franz.com/agraph/allegrograph/. Accessed Feb 6, 2015.

Gislason, H. (2013). BIG Interview by John Dominque.

Glavic, B. (2014). Big data provenance: Challenges and implications for benchmarking. In T. Rabl, M. Poess, C. Baru, & H.-A. Jacobsen (Eds.), *Specifying big data benchmarks* (pp. 72–80). Berlin: Springer.

Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006). Attribute-based encryption for fine-grained access control of encrypted data. In *Proceedings of the 13th ACM Conference on Computer and Communications Security* (pp. 89–98).

Grady, J. (2013). *Is enterprise cloud storage a good fit for your business?* http://www.1cloudroad.com/is-enterprise-cloud-storage-a-good-fit-for-your-business

Harizopoulos, S., Abadi, D. J., Madden, S., & Stonebraker, M. (2008). OLTP through the looking glass, and what we found there. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management data (SIGMOD '08)*.

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013). A discussion of privacy challenges in user profiling with big data techniques: The EEXCESS use case. *IEEE International Congress on Big Data* (pp. 25–30).

Hortonworks. (2014). Hortonworks Company Web Page. http://hortonworks.com/. Accessed Feb 6, 2015.

Ikeda, R., Park, H., & Widom, J. (2011). Provenance for generalized map and reduce workflows. In *Biennial Conference on Innovative. Data Systems Research* (pp. 273–283).

iQuartic. (2014). iQuartic pioneers use of Hadoop with Electronic Health Records (EHRs/EMRs). In: iQuartic Blog. http://www.iquartic.com/blog/. Accessed Feb 7, 2014.

Kamara, S., & Lauter, K. (2010). Cryptographic cloud storage. In *Financial Cryptography and Data Security* (pp. 136–149).

Kobielus, J. (2013) Big Data needs data virtualization. In: InfoWorld Webpage. http://www.infoworld.com/article/2611579/big-data/big-data-needs-data-virtualization.html. Accessed Nov 18, 2014.

Koomey, J. G. (2008). Worldwide electricity used in data centers. *Environmental Research Letters, 3*, 034008. doi:10.1088/1748-9326/3/3/034008.

Lee, C.-C., Chung, P.-S., & Hwang, M.-S. (2013). A survey on attribute-based encryption schemes of access. *International Journal of Network Security, 15*, 231–240.

Lee, K., Manning, S., Melton, J., Boyd, D., Grady, N., & Levin, O. (2014). Final SGBD report to JTC1, ISO/IEC JTC1 SGBD document no. N0095.

Li, M., Yu, S., Zheng, Y., Ren, K., Lou, W., et al. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. *IEEE Transactions on Parallel and Distributed Systems, 24*, 131–143.

Lobillo, F., Puente, M. A., Lippell, H., Zillner, S., Bretschneider, C., Oberkampf, H., et al. (2013). BIG deliverable D2.4.1 – First draft of sector's roadmaps.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Sydney: McKinsey Global Institute. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

MapR. (2014). MapR Company Website. https://www.mapr.com/. Accessed Feb 6, 2014.

Margaret Rouse. (2014a). *Object Storage*. http://searchstorage.techtarget.com/definition/object-storage. Accessed Nov 20, 2014.

Margaret Rouse. (2014b). *Block Storage*. http://searchstorage.techtarget.com/definition/block-storage. Accessed Feb 4, 2014.

Marz, N., & Warren, J. (2014). A new paradigm for Big Data. In N. Marz & J. Warren (Eds.), *Big Data: Principles and best practices of scalable real-time data systems*. Shelter Island, NY: Manning Publications.

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proceedings of the 18th International Conference Data Engineering*. IEEE Computer Society (pp. 117–128).

Miller, S. P., Neuman, B. C., Schiller, J. I., & Saltzer, J. H. (1987, December 21). *Section E.2.1: Kerberos authentication and authorization system*. MIT Project Athena, Cambridge, MA.

Mora, A. C., Chen, Y., Fuchs, A., Lane, A., Lu, R., Manadhata, P. et al. (2012). *Top Ten Big Data Security and Privacy Challenges*. Cloud Security Alliance (CSA).

Neo4j. (2015). Neo4j Company Website. http://neo4j.com/. Accessed Feb 6, 2015.

Okman, L., Gal-Oz, N., Gonen, Y., et al. (2011). Security issues in nosql databases. In *2011 I.E. 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (pp. 541–547).

OpenStack. (2014). OpenStack Website. http://www.openstack.org. Accessed Feb 3, 2014.

PEC. (2014). Peer Energy Cloud Project Website. http://www.peerenergycloud.de/. Accessed Feb 6, 2015.

Russel, J. (2013). *Cloudera Impala*. O#Reilly Media

Shenker, S., Stoica, I., Zaharia, M., & Xin, R. (2013). Shark: SQL and rich analytics at scale. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 13–24).

Sherif, A., Sohan, R., & Hopper, A. (2013). HadoopProv: Towards provenance as a first class citizen in MapReduce. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance*.

Shucheng, Y., Wang, C., Ren, K., & Lou, W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. *INFOCOM* (pp. 1–9).

Shvachko, K. H. K., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In *IEEE 26th Symposium on Mass Storage Systems and Technologies* (pp. 1–10).

Stonebraker, M. (2012a) What Does "Big Data" Mean? In: BLOG@ACM. http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext. Accessed Feb 5, 2015.

Stonebraker, M. (2012b). What Does "Big Data" Mean? (Part 2). In: BLOG@ACM. http://cacm.acm.org/blogs/blog-cacm/156102-what-does-big-data-mean-part-2/fulltext. Accessed Apr 25, 2013.

Strohbach, M., Ziekow, H., Gazis, V., & Akiva, N. (2011). Towards a big data analytics framework for IoT and smart city applications. In F. Xhafa & P. Papajorgji (Eds.), *Modeling and processing for next generation big data technologies with applications* (pp. 257–282). Berlin: Springer.

Thusoo, A., Sarma, J., Jain, N., Shao, Z., Chakka, P., Anthony, S., et al. (2009). Hive – a warehousing solution over a map-reduce framework. *Statistics and Operations Research Transactions, 2*, 1626–1629.

Turner, V., Gantz, J. F., Reinsel, D., & Minton, S. (2014). *The digital universe of opportunities: Rich data and the increasing value of the internet of things*. Framingham, MA: International Data Corporation (IDC). http://idcdocserv.com/1678. Accessed Aug 19, 2015.

Venkatesh, P., & Nirmala, S. (2012). NewSQL – The new way to handle big data. In: Blog. http://www.opensourceforu.com/2012/01/newsql-handle-big-data/. Accessed Nov 18, 2014.

Vogels, W. (2013). Day 2 Keynote. AWS reInvent.

VoltDB. (2014). VoltDB Company Website. http://www.voltdb.com. Accessed Nov 21, 2014.

Webber, J. (2013). BIG Interview by John Dominque.

White, T. (2012). *Hadoop: The Definitive Guide*. O'Reilly.

Wikipedia. (2013). *Column-oriented DBMS*. http://en.wikipedia.org/wiki/Column-oriented_DBMS. Accessed Apr 25, 2013.

Winder, D. (2012). Securing NoSQL applications: Best practises for big data security. *Computer Weekly*.

Wu, C., & Guo, Y. (2013). Enhanced user data privacy with pay-by-data model. In *Proceeding of 2013 I.E. International Conference on Big Data* (pp. 53–57).

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark : Cluster computing with working sets. In *HotCloud'10 Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing* (p. 10).

Ziekow, H., Goebel, C., Struker, J., & Jacobsen, H.-A. (2013). The potential of smart home sensors in forecasting household electricity demand. In *Smart Grid Communications (SmartGridComm), 2013 I.E. International Conferences* (pp. 229–234).

Zillner, S., Neururer, S., Munne, R., Prieto, E., Strohbach, M., van Kasteren, T., et al. (2014). D2.3.2. Final Version of the Sectorial Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).