IEEE Access
The journal for rapid open access publishing

Received September 2, 2015, accepted September 23, 2015, date of publication October 16, 2015, date of current version November 5, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2490723

Evaluating the Quality of Social Media Data in Big Data Architecture

ANNE IMMONEN, PEKKA PÄÄKKÖNEN, AND EILA OVASKA
VTT Technical Research Centre of Finland, Oulu 90571, Finland

Corresponding author: A. Immonen (anne.immonen@vtt.fi)

This work was supported by Tekes and VTT through the DIGILE's Need for Speed Program.

ABSTRACT The use of freely available online data is rapidly increasing, as companies have detected the possibilities and the value of these data in their businesses. In particular, data from social media are seen as interesting as they can, when properly treated, assist in achieving customer insight into business decision making. However, the unstructured and uncertain nature of this kind of big data presents a new kind of challenge: how to evaluate the quality of data and manage the value of data within a big data architecture? This paper contributes to addressing this challenge by introducing a new architectural solution to evaluate and manage the quality of social media data in each processing phase of the big data pipeline. The proposed solution improves business decision making by providing real-time, validated data for the user. The solution is validated with an industrial case example, in which the customer insight is extracted from social media data in order to determine the customer satisfaction regarding the quality of a product.

INDEX TERMS Architecture, big data, metadata, quality attribute, quality of data.

I. INTRODUCTION

Nowadays there is a lot of freely accessible data available online. This data is made available by different parties, such as public sectors, private companies, different organizations and institutes, single individuals and the different forms of social media. As the amount of data is enormous, the term 'big data' becomes apparent, meaning a massive volume of structured and/or unstructured data being too difficult to process using traditional database and software techniques. Benefits of open data [1] have already been discovered widely around the world. Several public sectors and even private companies have been interested in opening their data, as data exploitation has been recognized to include several benefits for businesses [2]. Recently, also social media data, such as data from Twitter and Facebook, has increasingly interested companies in their business decision making, as these free-formed discussions can provide insight into consumers' opinions, preferences and requirements considering the company or its products/services [3]–[5]. Big Data Initiatives already exist, spreading out in all directions and comprising various themes, tending to end up in innovative economic development. For example, there are political initiatives, like Big Data – Big Deal,[1] promoted by the Whitehouse. A European initiative[2] by the Big Data Value Association focuses on creating value of big data, whereas NIST[3] and researchers in computer science advanced education in India[4] introduce R&D Initiatives. The terms of 'open data' and 'big data' have been familiar concepts also for many companies for several years. At this moment, a new challenge for companies is to develop a business model around these concepts and create new value from the data through large-scale analytics [6]. The big data dimensions; volume, variety, velocity and veracity [6], pose challenges not only to data analytics, but also to the big data systems that must manage all the data.

As a lot of freely accessible data is commonly unstructured or not more than semi-structured [7], [8] and originates from indeterminate sources, the quality and trustworthiness of the data become key issues. Data quality can be defined according to [9]; data that are fit for use by data consumers. Trustworthiness of data has a broader meaning, defining the perceived likelihood that a piece of information will preserve a user's trust in it [10], and consisting of factors that influence how data-users make decisions regarding the trust

[1] https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

[2] http://www.bdva.eu/

[3] http://www.nist.gov/itl/ssd/is/upload/NIST-BD-Platforms-05-Big-Data-Wactlar-slides.pdf

[4] http://drona.csa.iisc.ernet.in/~bigdata/

in information. The data-users (in this case, the companies) need to ensure the quality and trustworthiness of data and be able to trust in it in their businesses. At first, when collecting data, the user wants to ensure the reliability of the data and the data source, leaving out suspicious data. Secondly, when further processing and analyzing the data, the user wants to ensure that the quality and relevancy of data are appropriate for the specific situation. Reliable and valuable data enhances business decision making in several ways, enabling, for example, real-time demand predictions, the estimation of trends, and innovation of potential new products/services. The usage of unreliable data, such as data from suspicious sources, or corrupted, subjective, inaccurate or incomplete data, has a high risk for a company's business, and may lead to poor or incorrect business decisions. Furthermore, the usage of valueless and irrelevant data for certain situations causes a lot of unnecessary effort and expenses for companies.

The evaluation of data quality has relevance in one or more data processing phase(s) of big data architecture (i.e. big data pipeline); in data extraction, data processing and analysis, and finally in decision making. Therefore, quality evaluation of big data must be considered during architecture design, when designing how the data goes through the pipeline of a big data system. Difficulties in quality evaluation are determined by the fact that data quality cannot be judged without considering the context at hand [10]; the same quality attribute is applicable to different situations but the evaluation metric is different. In addition, there are no agreed definitions of quality attributes or classification of their applicability to certain contexts. Furthermore, the characteristics of big data, [6], [11], and [12] as such, set special challenges for quality evaluation. The growing amount of semi-structured and unstructured data, new ways of delivering information and user's changed expectations and perceptions of data quality have been recognized as new challenges in data quality research [8]. Thus, it is obvious that new means are required for data quality evaluation for such kinds of big data.

The purpose of this paper is to describe how to ensure the quality and trustworthiness of social media data for company's business decision making. We introduce a novel solution for data evaluation, in which the data consumer can select the applicable quality attributes and evaluation metrics for the context and situation at hand, and evaluate the quality attributes with evaluation metrics. The solution follows the pipeline of the big data reference architecture of [7].

This paper is organized according to the following: Section 2 defines the basic terms used in this work, and provides state-of-the-art of the big data architectures, and the application of metadata, quality attributes, quality metrics and quality policies in business usage. Section 3 introduces our solution for data quality evaluation in big data architecture. Section 4 provides a case example of how the developments are used in practice; an industrial case company achieves insight into customer needs utilizing social media data. Section 5 provides the validation of the trial usage of the solution and identifies the shortcomings and development targets. Finally, section 6 concludes the work.

II. BACKGROUND

A. TERMINOLOGY

The following terminology is used in this paper:

Data – Data that is produced by observing, monitoring, or using questionnaires, but has not yet been processed for any specific purpose.

Big data – Data that is numerous, cannot be categorized into regular relational databases, and is generated, captured, and processed rapidly [11].

Big data architecture – An architecture that provides the framework for reasoning with all forms of data [13]. Thus, it is a logical structure of core elements used to store, access and manage the big data.

Information – Data that is refined and processed for assigning meaning to the data [14].

Knowledge – Understanding of a subject. Knowledge can be implicit or explicit, and it is more or less systematic. Theoretical knowledge represents explicit knowledge on the meaning of data. Practical knowledge is implicit and less systematically collected, represented and shared.

Service – A digital service that provides additional value for data processing and can, for example, support data collection, analysis, sharing and/or representation [2].

Quality attribute – A representation of a single aspect or construct of a quality [9].

Quality metric – A measure of certain properties of the quality attribute, evaluating the degree of presence of the quality attribute [15].

Quality assessment – Assessing of the quality of raw data as such, without considering the context or the intended use of data.

Quality evaluation – Evaluating the quality of information, taking into account the context and the intended use of information.

Metadata – Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource [16]. Quality metadata describes the quality attributes of the data and the metrics for each quality attribute.

Quality policy – A policy is a collection of alternative tasks and rules, each of them representing a requirement, capability, or other property of behavior [17]. Quality policies are used to generate quality objectives, serving also as a general framework for action [18].

B. BIG DATA ARCHITECTURES

Big data can be categorized according to data sources, content format, data stores, data staging and data processing [11]. Each of these categories represents several new challenges to data-intensive systems. To achieve high performance, availability and scalability, the big data systems are often distributed. Both software and data architecture must be resilient; the data must be replicated to ensure

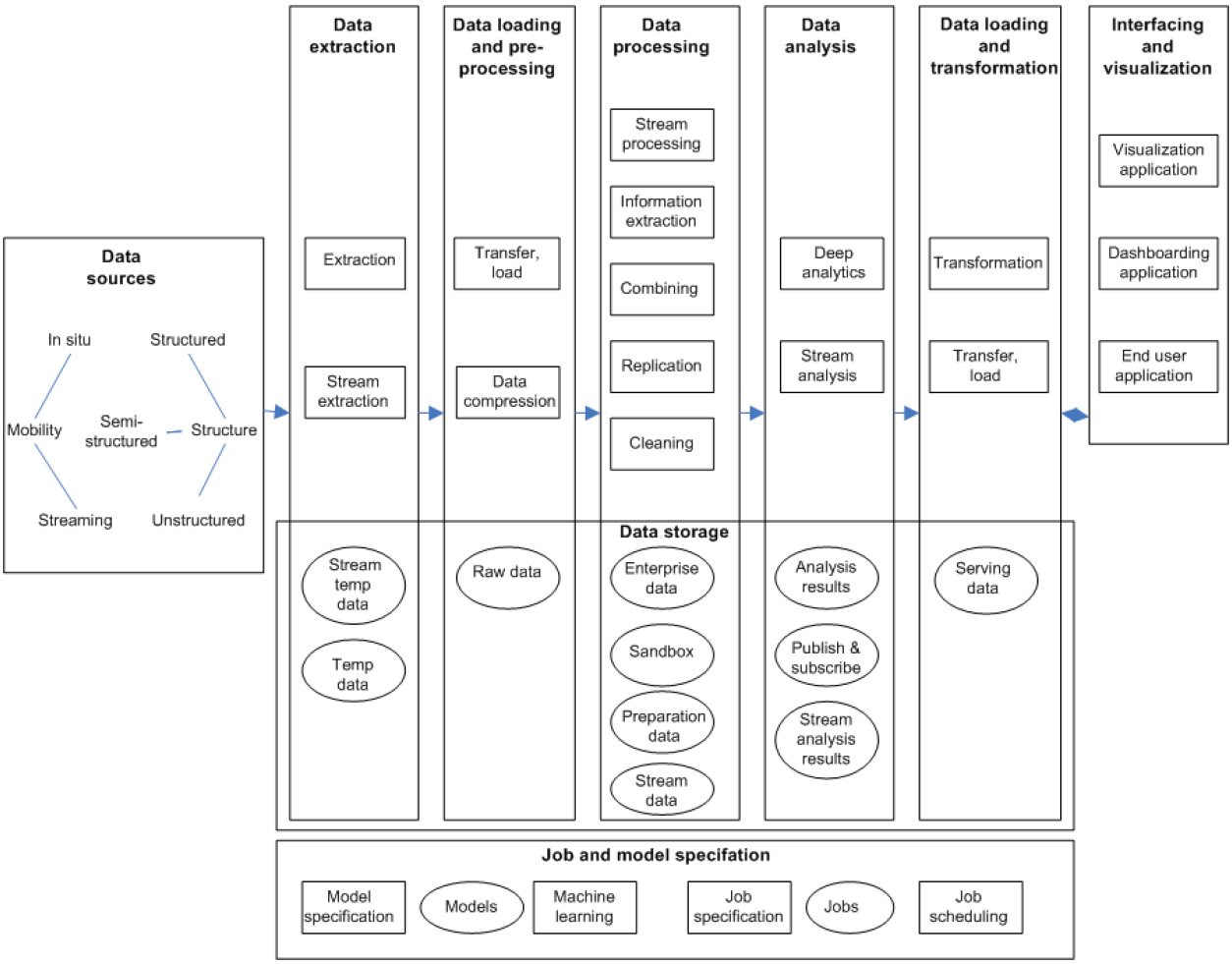

FIGURE 1. High-level view of big data reference architecture [7].

availability and the components of the architecture must be stateless, replicated and failure tolerant [19].

Several implementation architectures of big data systems have been published based on commercial services (Facebook, Twitter, LinkedIn, Netflix, etc.). Recently, a big data reference architecture [7] was published, which had been missing from earlier literature. The big data reference architecture is based on the analysis of published implementation architectures of big data systems. Fig. 1 describes the high-level design of the reference architecture (see [7] for a detailed description and related technologies) derived from published big data use cases. The architecture consists of functionalities (depicted with a rectangle), data stores (circles), and data flows (arrows) between them. Data flows typically from left to right in a big data pipeline. In a big data system, data may be extracted from different sources and stored in a temporary data store. Data may also be further loaded and transmitted into a raw data store, and processed for extraction of new information (to be stored into an enterprise data store). Further, the gathered data is typically analyzed, and results are stored (into a data analysis store). Finally, the analyzed results may be further transformed for serving applications and visualization purposes.

The reference architecture does not consider metadata aspects of big data, which are focused on in this paper.

C. METADATA AND METADATA STANDARDS

The properties of data, such as provenance, quality, and technical details, can be described in metadata of the data, which is simply 'data about data'. Thus, metadata assists end-users to validate the quality and value of data for business usage. However, at this moment the end-users are only slightly satisfied with the metadata available to them [20], and the recent metadata standards do not assist in finding out the quality of data from the data end-user's viewpoint.

Metadata is commonly classified in three categories: descriptive, structural, and administrative metadata [16]. Descriptive metadata identifies a resource and describes its intellectual content. Structural metadata indicates how compound objects are put together, supporting the intended presentation, and use and navigation of a data object. Administrative metadata provides information necessary to

allow a repository to manage objects, such as when, how and by whom a resource was created and how it can be accessed. Metadata standards intend to establish a common understanding of the meaning or semantics of the data. A lot of work has been done by international standardization bodies on standardizing metadata and registries [16], [21]. Data exchange between systems is accomplished by using architectural principles of computer and software systems. The Common Warehouse Metamodel (CWM) [22] is a de-facto standard for data integration by specifying metadata for different kinds of objects found in a data warehousing environment. ISO/IEC 11179 [23] is a standard for metadata-driven exchange of data in a heterogeneous environment, defining metadata and activities needed to manage data elements in a registry. Moreover, the Dublin Core metadata element set enables service creators to describe their own Web resources [24].

A study among data end-users reveals that the end-users consider data quality metadata to be the most useful in metadata [20]. Although several metadata standards exist, it is difficult to estimate their advantages and choose the most applicable one. Furthermore, the standards do not consider data quality aspects from the data users' viewpoint. A data-user metadata taxonomy suggested by [20] facilitates the understanding of various information resources. The taxonomy includes four classes:

- Definitional metadata describes the meaning of data from a business perspective.
- Data quality metadata describes the quality of data when using it for a specific purpose.
- Navigational metadata helps users find the desired data.
- Lineage metadata describes the original source of data and the actions on the data.

D. QUALITY ATTRIBUTES AND METRICS

Several classifications of data quality attributes exist in the literature, but although almost 200 terms for data quality exist, there is no agreement regarding their nature. Some of the quality attributes are too abstract and lack agreed upon specifications for concepts and/or metrics for their evaluation. A lot of work has been done in standardizing quality attributes in the field of software engineering [15], [25], [26]. Quality has also been taken into account systematically in many works dealing with software architecture design [19], [27]–[31]. However, in the case of data, quality issues are not commonly brought into use. Data quality attributes have traditionally been classified into four dimensions important to data consumers [9]. The intrinsic dimension denotes that data have quality in their own right that is independent of the user's context. The contextual dimension considers quality within the context of the task at hand and the subjective preferences of the user. The representational dimension captures aspects relating to information representation, whereas the accessibility dimension captures aspects involved in accessing information. Several other works on data quality and trustworthiness attributes exist, such as [32]–[35], some of them even focusing on social media [36], [37]. The recent research on the quality of online data has been reviewed and summarized under three main factors [10];

- Provenance factors refer to the source of information.
- Quality factors concentrate on factors that reflect how an information object fits for use.
- Trustworthiness factors influence how end-users make decisions regarding the trust of information.

The quality metrics are often designed in an ad-hoc manner to fit a specific assessment situation [38]. Quality assessment metrics can be classified into three categories according to the type of information that is used as quality indicator [38]. Content-based metrics use information to be assessed per se as quality indicator, whereas context-based metrics employ meta-information about the information content and the circumstances in which information was created or used as quality indicator. Rating-based metrics rely on explicit ratings about information itself, information sources, or information providers.

E. QUALITY POLICIES

The quality policy defines which quality attributes are relevant in the context of the task at hand, which quality metric should be used to evaluate the defined quality attributes, and how the evaluation results should be compiled into an overall decision of whether to accept or reject information [9]. A company's organizational policy describes the principles and guidelines required to effectively manage and exploit data/information resources, whereas decision making policy is required for configuring quality evaluation according to the needs of the data-consumer.

The importance of quality policies has been recognized in several works. The Information Quality Assessment Framework [39] enables information consumers to apply a wide range of policies to filter information. The filtering policy consists of a set of metrics for evaluating the relevant quality dimensions, and a decision function that aggregates the resulting evaluation scores into an overall decision on whether information satisfies the information consumer's quality requirements. The approach described in [40] uses an information source filter for subscribing to a set of known information sources, and a scoring function to capture the provenance factors of interest and to assign scores to messages for each factor. The decision making policy allows the decision maker to amplify or attenuate one or more provenance factors that may appear to be more or less important in a particular situation. The framework proposed in [41] uses policies to specify the confidence level required for use of certain data in certain tasks, consisting of three major components: trustworthiness assessment, query and policy evaluation, and data quality management.

Although some promising policy-based approaches already exist for quality evaluation [39]–[41], their practical application is missing. In this work, the represented data quality evaluation solution applies the quality policies.

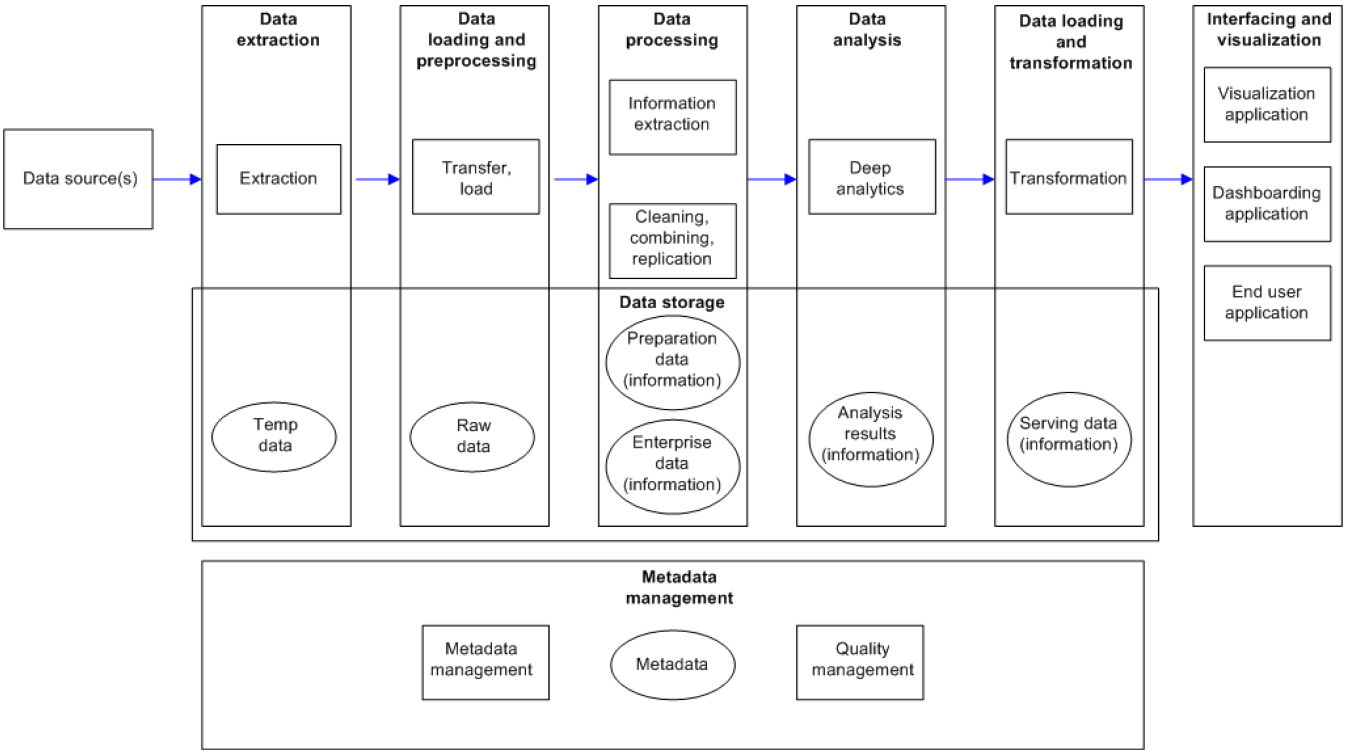

FIGURE 2. Metadata management in big data architecture (enhanced from Fig 1).

III. QUALITY EVALUATION IN BIG DATA ARCHITECTURE

The main purpose of our solution is to evaluate the quality and trustworthiness of data, and incorporate the valuable analyzed results of the data into a company's business decision making process. Data evaluation is conducted in several data processing phases of the big data architecture, going through the pipeline of a big data system. The elements and main phases of the approach are described in the following sub-sections. The metadata in our solution consists of several metadata groups; the whole metadata set is described in the next subsection, but since our focus is on quality, the rest of the paper concentrates only on the quality viewpoint.

Our solution utilizes the big data reference architecture of [7], adding the metadata management element into the big data pipeline (Fig. 2). The metadata management consists of one data store; metadata, and two functionalities: metadata management and quality management. 'Metadata' is a data storage used to store, organize and manage the metadata. 'Metadata management' enables extraction of metadata, and access to metadata. 'Quality management' assigns values to quality attributes based on the properties of associated metadata and data sets.

A. DATA AND METADATA IN THE BIG DATA PIPELINE

1) DATA AND DATA REFINEMENT

Fig. 2 describes the flow of data and creation of information through the big data pipeline. The data that is extracted into a big data system may be structured, semi-structured, or unstructured. Structured data has a strict data model (e.g. based on a database schema). Semi-structured data is not raw data or strictly typed, but instead it has an evolving data model (e.g. JSON/XML documents) [7]. Unstructured data is not associated with a data model, and can have miscellaneous content, such as documents, pictures, videos, etc. Data is extracted from the data source to a company's system as a data set that is an identified collection of data that contains individual data units organized in a specific way and collected for a specific purpose. Extracted data may be stored temporarily (into temp data storage), until it is loaded and/or preprocessed, and stored permanently to raw data storage.

When the data is processed, i.e. cleaned, replicated, combined or compressed, the raw data is transformed to enhanced data, and stored temporarily into preparation data storage. New information may also be extracted from raw data, and saved into enterprise data storage (by storing raw data in a structured format [7]). Deep analytics creates additional insight based on data/information, and entirely new data sets may be created in the form of analysis results. The analysis enables getting value from data and increasing data consumer's understanding of the data; thus transforming the data into information. Data transformation finally modifies analysis results for serving end-user applications (e.g. servicing of analytical queries).

2) METADATA GROUPS

In our approach, metadata is defined as data about gathered data sets in a big data pipeline. The metadata of the data set

TABLE 1. Quality attributes of quality metadata.

Quality attribute	Description and rationale
Accuracy	The degree of correctness and precision. Ensures that the data/information is error-free and the value is in consistent form in accordance with the business data model.
Believability	The extent to which information is regarded as true and credible. When the identity of the informer is known, the information is assumed to be more reliable, traceable, and less likely malicious.
Completeness	The degree to which data/information is not missing. Verifies that the data/information is sufficient in breadth, depth and scope.
Consistency	Implies that two or more values do not conflict with each other. Ensures internal validity.
Corroboration	The same data comes from different sources. Freely available online data can be assumed to be true when the same data comes from several different sources.
Coverage/ amount of data	The extent to which the volume of data is appropriate for the task at hand (appropriate volume of data available). This means that information is of sufficient breadth and depth for the task of the information consumer. The coverage (breadth and depth) can be assessed for each data set, and the large amount of data sets provides assurance in decision making
Validity	Indicates the likelihood that the information is valid in a certain situation.
Popularity	The source provides accurate information, having a number of followers, or the information is liked and therefore repeated by others. The widespread use of a resource tends to lead to more trust.
Relevancy	The extent to which information is applicable and helpful for the task at hand. Non-relevant data sets should not be considered further.
Timeliness	The freshness of the data; timestamp is important for extracted, processed and analyzed data sets.
Verifiability	The degree and ease with which the data/information can be checked for correctness. The traceability and provability of data/information; the data can be verified by users, for example, by using the references to original sources.

is divided into five groups based on the existing standards (e.g. [23] and [24]);

1. Navigational metadata (i.e. where the data set can be found) provides the list of semantic tags or keywords identifying the data set, and the location where the data set can be found.
2. Process metadata (i.e. where did the data originate from and what has been done to it) describes the original source of data, processing performed on the data set and the processing application.
3. Descriptive metadata (i.e. what does the data mean) consists of business and technical metadata. The business metadata describes the meaning of the data set from a business perspective (e.g. a link to the organizational policy to be used in evaluation of the data set) and its purpose for decision making (e.g. a link to the decision making policy to be used in evaluation of the data set). The technical metadata provides the technical information of the data set, such as a unique identifier, the language and size of the data, content description, data creator and creation place, content type and format, and required software to render and use the data.
4. Quality metadata (i.e. the applicability of the quality of data for its intended use) consists of the attributes (e.g. timeliness and accuracy) and the metrics that describe the quality of data.
5. Administrative metadata (i.e. how to access and use the data) describes the data provider, the applicable license(s) and access rights on the data set, the copyright holder and indicator of the data privacy level.

This work concentrates on quality metadata, assuming that other groups of metadata also exist.

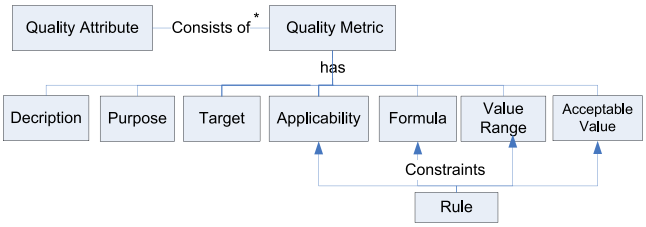

FIGURE 3. Properties of data quality metrics (adapted from [42]).

3) DESCRIPTION OF QUALITY METADATA

Table 1 describes the attributes of quality metadata. Each quality attribute describes a single aspect or construct of a quality. A quality attribute consists of one to several quality metrics that are measures of certain properties of the quality attribute (**Fig. 3**). Each metric has the following properties (adapted from [42]):

- Description: the description of the metric
- Purpose: the description of the metric purposes
- Target: where the metric can be used.
- Applicability: when the metric can be used.
- Formula: how the value for the metric is achieved.
- Value range: the range value for the metric measurement/evaluation.
- Acceptable value: the minimum measure accepted for the quality attribute.
- Rule: the set of constraints defining the set of targets of measurement, the set of value ranges for the measurement unit and the time when the metric is valid.

4) SELECTING QUALITY METADATA ATTRIBUTES FOR A DATA SET

The collected data can be of different types, such as a) any freely available data according to a company's interests,

TABLE 2. The application of social media data quality attributes.

Quality attribute	Applicability time	Metric (examples for twitter data)
Believability	Extraction, analysis	Evaluating the believability of a source (e.g. the identity of the authors): • Registration age = the time passed since the author registered his/her account, in days • Statuses_count = the number of tweets/comments/ questions at posting time • Followers_count = the number of people following this author • Friends_count = the number of people this author is following • Is verified = the author has a 'verified' account
Corroboration	Analysis	The amount of analyzed data sets from which the identified issue is recognized
Validity	Extraction, analysis	Estimation of the likelihood (0…1) of data validity in its purposed usage
Popularity	Extraction	• Popularity of source = the number of readers, followers, etc. • Popularity of information = the number of re-tweets, comments, questions, etc.
Relevancy	Extraction, processing, analysis	The amount of occurrence of relevant key words in title, subject and description
Timeliness	Extraction, processing, analysis	The data set creation date

originating from uncertain sources from the Internet, (e.g. data from web pages or from social media), b) deliberately collected external data from reliable or uncertain sources for certain internal process purposes, (e.g. for market analysis and competitor analysis), c) customer feedback data that can be reliable or uncertain, depending on the way the feedback is given, or d) a company's internal data, such as product data and production data. The collected data is classified according to data source types, such as social media data, feedback data, product data, competitor data, history data, or production data. This classification assists in selection of applicable quality attributes for the metadata of the given data set. The attributes are classified for each data source type. For example, the attributes applicable for social media data are described in Table 2. Thus, for example, each data set with the data source type ''social media data'' has quality metadata with the same quality attributes in a specific situation.

B. DATA QUALITY EVALUATION IN DATA PIPELINE

1) EVALUATION PHASES

In our approach, the metadata is managed in the following phases: data extraction, data processing, data analysis, and decision making. The first three phases follow the big data pipeline (Fig. 2). In the decision making phase, the analyzed data is visualized to the data user with varying views and varying users controls (Interfacing and visualization in Fig. 2); without user control, limited set of control functions or detailed visualization and control functions. The decision making based on the visualized data is the responsibility of the data user (according to decision making policies of the company).

Fig. 4 describes the different evaluation focuses and viewpoints on data. In data extraction, the focus is on the data source, when quality evaluation focuses on data provenance and the data quality from the viewpoint of the situation at hand, i.e., why the data was extracted. In data processing and analysis the focus is on data itself, evaluating the different quality aspects of the data. In decision making, the data is examined from the data user's viewpoint (i.e. data in context), when the evaluation focuses on data trustworthiness, i.e., how to ensure the trustworthiness of data in decision making. The data provenance and data quality assists in trustworthiness evaluation.

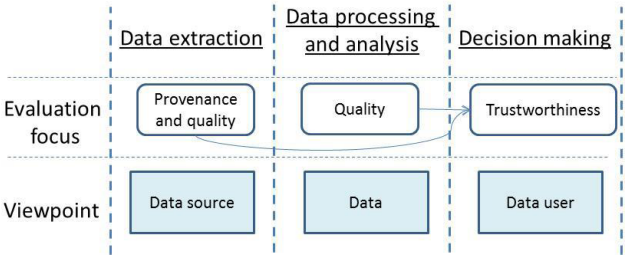

FIGURE 4. The focuses and viewpoints of data quality evaluation in the metadata management phases.

2) EVALUATION OF QUALITY ATTRIBUTES

The evaluation of quality attributes occurs in each metadata management phase. Quality evaluation can be qualitative or quantitative. Quantitative evaluation is a systematic and formal process, applicable in all metadata management phases. It relies on the existing knowledge of the company defined by the rules via a company's quality policies (see Section III C), and applies computational methods to achieve values for the metrics. The results of the quantitative evaluation are objective and more concrete than in the case of a qualitative evaluation. The quantitative evaluation can be automatized, and it can be performed by the company itself or it can be outsourced to third-party evaluation service providers.

Qualitative evaluation relies on the existing knowledge of the company, and also the experience and knowledge of the evaluator (expert or professional). The qualitative evaluation is applicable in data extraction, when the purpose of the data extraction is linked from business metadata to a company's

quality policies, and in decision making, when the value of the data is evaluated in the context of the current situation.

C. QUALITY METADATA MANAGEMENT IN BIG DATA SYSTEMS

To manage quality metadata, attributes and metrics, rules (see Fig. 3) are needed to define variability in quality, i.e. which quality attributes and metrics can be used and when. The rules can be described, for example, by a simple if-then-else structure or using some rule language, such as [43] and [44]. These rules should be part of a company's quality policy, which defines the principles and guidelines on how to manage quality in the company. The quality variability and quality policies are described in the next sub-sections.

1) DESCRIPTION OF QUALITY VARIABILITY

Different types of variation among quality attributes exist that describe a data set:

1. Target of attribute: Certain quality attributes are applicable only to certain data source types. For example, believability is an important attribute for data of which the origin is unclear. However, believability of a company's production data can be assumed to be high; thus the believability attribute is irrelevant. The quality attributes are selected based on the source type of the data set.
2. Applicability of attribute: Some of the quality attributes are applicable in the data extraction, some in the data processing or analysis, whereas some are applicable in all three phases. For example, corroboration cannot be evaluated for a single data set in data extraction, but it is important when evaluating several data sets in the analysis phase.
3. Target of metric: There are different metrics that can be used to evaluate a quality attribute. The selection of the metric is dependent on the data source type. For example, a different metric is used to evaluate corroboration in the case of twitter data or in the case of feedback data.
4. Applicability of metric: Different metrics can be used to evaluate the attribute in different phases. For example, the coverage of data is evaluated in data extraction phase by inspecting the amount and the content of data of the single data set, but in the analysis phase the coverage can be defined simply by the amount of the data sets

2) DATA QUALITY POLICIES

The data sets and metadata are administrated by the company's quality policies. The terms organizational policy and decision making policy have been adopted from [40]. The organizational policy defines the acceptable data sources, and describes all the elements from Fig. 3, such as the relevant quality attributes applicable to the context of the task at hand, the applicability time of the attributes, which evaluation metric should be used to evaluate each attribute, etc. Thus, the organizational policy consists of the set of rules that describe what and how to evaluate to achieve the data that can be trusted in a specific situation. A company can have several organizational policies, each of them applicable for different purposes of data collection.

The decision making policy describes which data sets are relevant for certain situations, how to weight quality attributes depending on the relevance of the different quality attributes for the task at hand, and how to perform the decision functions. The company can have several decision making policies, each of them describing the rules of how to make decisions in certain situations. Each policy can be applicable to different purposes of data collection/analysis or for different stakeholders. In addition, decisions are made during different stages of the product/service development process: in pre-development, development and post-development [5]. In the pre-development stage, the collected data is used in requirements specifications. During development, the data is used to identify modifications for the product/service and is an important input for further improvement. Finally, in the post-deployment stage, the data is used to optimize or innovate new features for a current or new product. The selection of the appropriate decision making policy is based on the existing experiences and knowledge of the company.

Both the organizational policy and decision making policy must be configurable by the data user to adapt the policies to the situation at hand. The user should be able, for example, to define the acceptable data sources, add new data sources, add new metrics/methods and configure the acceptable values for the quality metrics according to the context and purpose for the data collection. In the same way, during decision making, the user may want to configure acceptable values for the quality attributes for data set selection for decision making, or weighing quality attributes according to a new, changed situation.

D. PERFORMING THE DATA QUALITY EVALUATION AND MANAGEMENT IN BIG DATA SYSTEMS

This section describes how the previously introduced elements are used in the different evaluation phases in the big data pipeline, and what architectural elements are needed to enable data quality evaluation and management.

1) USING THE SOLUTION FOR QUALITY EVALUATION OF EXTRACTED DATA

Fig. 5 describes the activities that the end-user performs when using the solution for data extraction. These activities can be modified to be applicable also in the case of data processing and analysis. The main rationale for data collection is to assist a company in business and decision making; therefore, the meaning and purpose of the data collection must be defined beforehand (step 1 in Fig. 5). The purpose is later added into the descriptive business metadata (see section III A2 bullet 3). The metadata facilitates managing the data sets and enables the users to validate the value of data. The metadata is managed by the organizational and decision making policies,

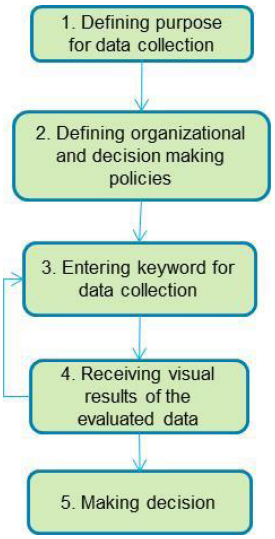

FIGURE 5. The user activities for data quality evaluation.

which must be defined applicable to certain business goals or certain types of purposes (Step 2 in Fig. 5). After that, the data and metadata management are automatically guided by the policies. Step 1 and Step 2 should be carefully defined, since they describe the reason and rules for data collection and evaluation. The end-user can collect data, for example, by entering a search keyword through the user interface (Step 3 in Fig. 5). The solution automatically extracts the data, evaluates the data quality, and finally visualizes it to the user according to the quality policies (Step 4). After seeing the results, the user may want to change metadata values (going back to the Step 3) and bring more data sets into the evaluation. Finally, the end-user makes business decisions based on the data (Step 5).

Our solution enables automatic data quality evaluation and management. Quantitative evaluation can be entirely automatized. Since the qualitative evaluation is managed mainly by human experts or professionals, it requires visualization of metadata to the user, and a user interface that enables the user to input values into the metadata (adding a new step between the steps 3 and 4).

2) CREATION OF QUALITY METADATA IN THE BIG DATA PIPELINE

Fig. 6 represents the data extraction, processing, analysis and decision making functionalities as an activity diagram. The functionalities are assisted by quality policies, in which the company's knowledge is presented by rules. In data extraction, the organizational policy facilitates the process by defining the acceptable data sources, and in selection of acceptable quality attributes, applicability time of the quality attribute and metrics and methods to evaluate the quality attributes. The applicable attributes are automatically provided when the data source type of the data set is known. The quality attributes are then evaluated using qualitative and/or qualitative evaluation. After extraction the imported data is stored in data storage. The quality metadata is created for the data set and the evaluated values for quality attributes are automatically inserted into the metadata. The metadata is stored in a metadata registry, separately from the data set.

In the same way, the organizational policy helps to select data sets for processing/analysis purposes. For example, the policy can set the value range for the quality attributes in metadata; only the data sets whose evaluated quality attributes fulfill the policy requirements defined for the processing/analysis phase are accepted, others are discarded. The organizational policy also assists in attaching the applicable quality attributes for the metadata of the data set and the metrics in this phase for evaluating the quality. After evaluation the quality metadata is created for the processed/analyzed data set and the evaluated values for quality attributes are inserted into the metadata.

During decision making, the decision making policy facilitates the selection of relevant data for the decision making purposes, e.g., by defining the important quality attributes and the minimum values for the selected data sets. That is, the policy defines which data sets are important for the situation at hand, and also validates their reliability and value for decision making. When evaluating the significance of a data set for a certain purpose, the decision making policy helps to weight the relevant quality attributes for the particular situation. The data is visualized to the data user with a visualization application with certain views and controls on data (defined in decision making policy). Decision making policy is always dependent on the company, its priorities and the goals and purposes for data gathering and analysis.

3) ARCHITECTURE FOR METADATA MANAGEMENT AND QUALITY MANAGEMENT

The architecture for the data, metadata management and quality management includes several elements of Fig. 2 with the following responsibilities:

- Extractor; extracts the data from data sources
- Temp data store; stores the extracted data temporarily
- Deep analytics; performs batch processing-based analysis for the collected data sets
- Analysis results store; stores the analysis results permanently
- Metadata management; responsible for creating, updating, storing and accessing the metadata
- Quality management; manages quality metadata for the data sets utilizing the company's quality policies and quality evaluator services. It includes the following complementary elements: Quality policy manager for the management of a company's quality policies, and Quality evaluator for evaluating the values for the metrics of the quality attributes.
- Metadata store; stores the metadata of the extracted, processed and analyzed data sets
- End-user application; provides the user interface to manage the data extraction, processing and analysis,

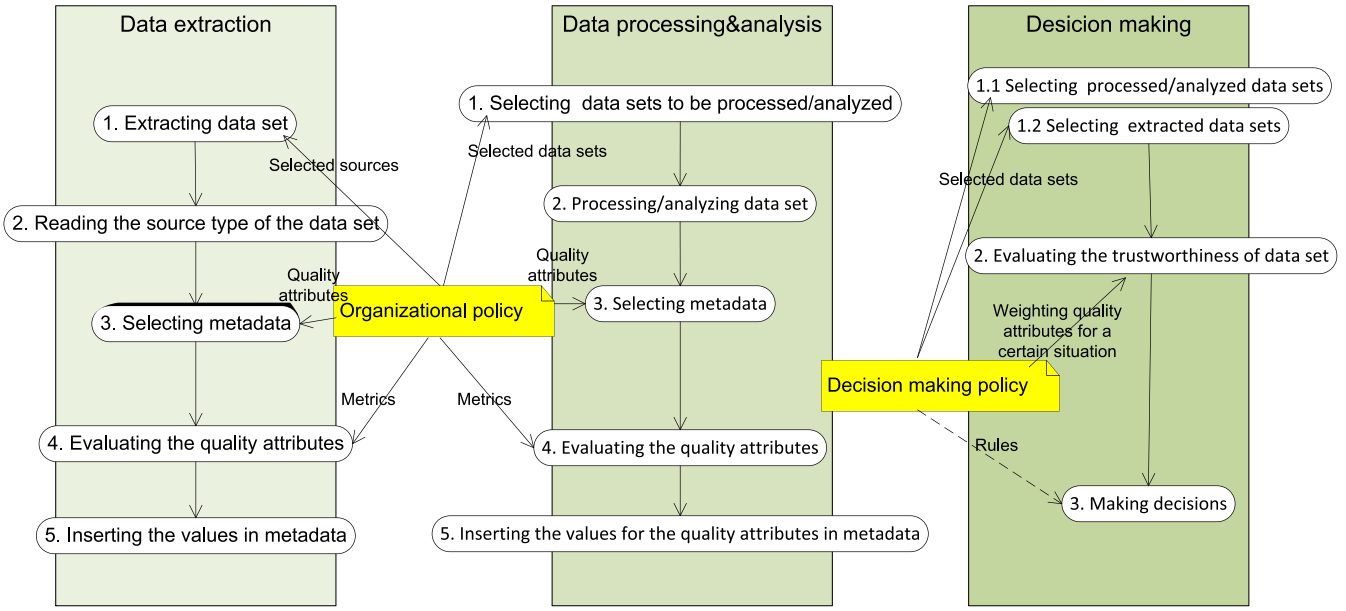

FIGURE 6. **Creation of quality metadata in different phases of metadata management in the data pipeline.**

enables the end-user to configure the quality policies, and visualizes the analysis results for the end-user.

Fig. 7 describes the architecture of the new element of Fig. 2; Metadata management, in more detail. Metadata management enables organization and management of the metadata of the data sets, and also the creation and management of quality metadata, enabling data quality evaluation. The architectural elements and their responsibilities are described in Table 3.

IV. CASE EXAMPLE

We demonstrate our solution using an industrial case example; the solution provides to the case company (a big data consulting company) insight regarding customer needs, which may facilitate R&D of the company. The data for the case example has been gathered by interviewing the case company's representatives. Also, the case example was implemented together with the case company, who wants to utilize social media data to find out what is discussed about their customers' products. The main purpose of the company is eventually to combine social media data with the company's own, internal data to achieve 'customer insight' that can be utilized in business decision making. The organizational and decision making policies have a great importance in quality evaluation; the definition of these policies is an organizational issue and is required as prerequisites for using the solution.

Fig. 8 describes an instantiation of elements in Fig. 2, illustrating the steps of data management in the case example at the architectural level.

Step 1 (Data Extraction and Analysis): At the data extraction phase, the end-user searches for relevant data using keywords. The keywords may be related, for example, to customers' products, and they are used for extraction of related tweets from Twitter. The tweets are extracted and saved into a temporary data store, and finally the sentiment of the tweets is analyzed. The case company has to define the acceptable data sources by the organizational policy before extraction of tweets. Step 1 of Fig. 8 is described in in more detailed in the following:

1.1 The end-user specifies keywords related to interesting commercial products.
1.2 DataExtractor extracts tweets via Twitter API based on the specified keywords (with a HTTP GET).
1.3 The tweets are saved into a temporary data store.
1.4 Deep analytics fetches the stored data sets from the TempData store after a certain time period.
1.5 Deep analytics performs sentiment analysis on the data sets. The aspect-based sentiment analysis [45] is used to analyze the sentiment of each individual aspect (words) in the discussion about the product, and to provide a sentiment score for the whole discussion.
1.6 The analysis results are saved into the analysis results store.

Step 2 (Metadata Creation and Data Quality Evaluation): This step focuses on creation of metadata in the big data pipeline. The metadata and related quality attributes are created based on the data sets of tweets (created in step 1) and the attributes are evaluated. The navigational, process, descriptive and administrative metadata are also created, but are not focused on in this paper. Step 2 of Fig. 8 is described in in more detailed in the following:

2.1 After saving the analysis results, Deep analytics notifies Metadata management to create metadata for

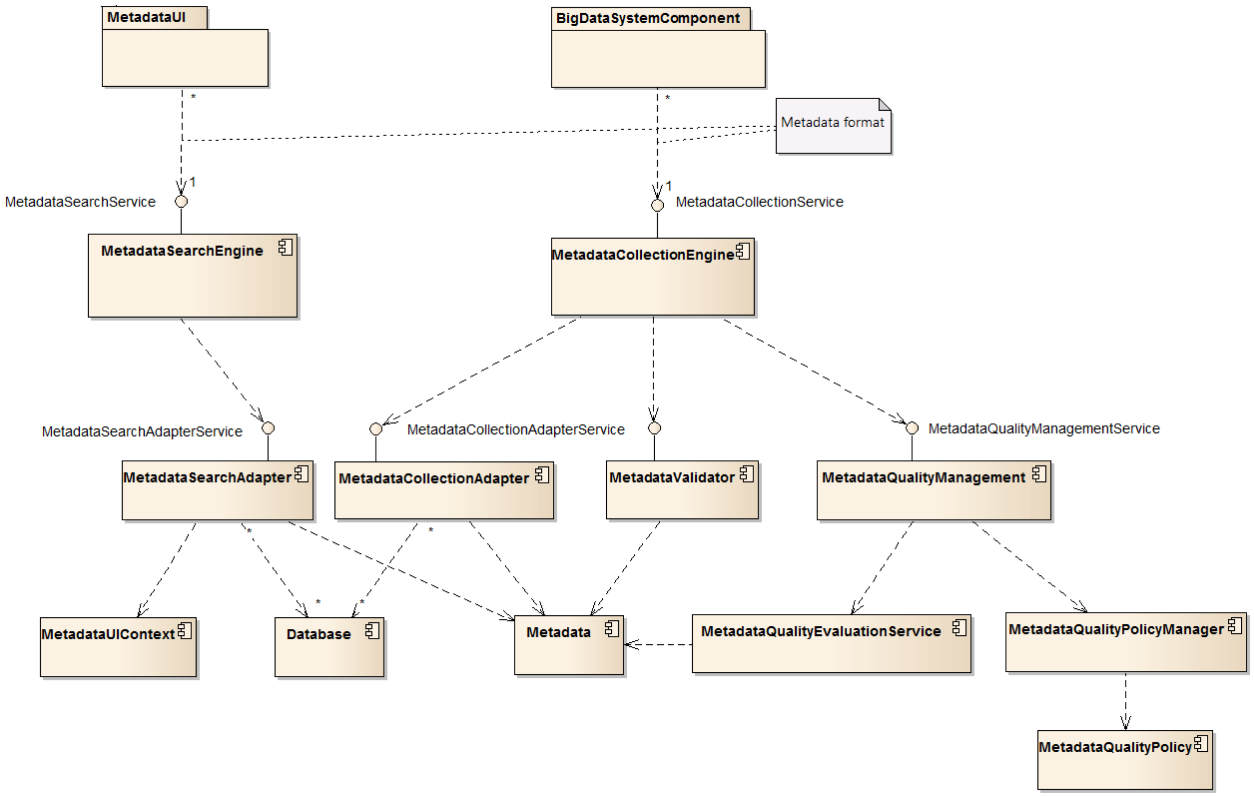

FIGURE 7. Structural view of metadata management architecture.

TABLE 3. Architectural elements of metadata management.

Element	Responsibility
MetadataCollectionEngine	Enables the extraction of metadata in a big data system via MetadataCollectionService API. The API enables external components of a big data system to input metadata to the Metadata store.
MetadataSearchEngine	Enables the searching for external components (e.g. through an UI) for metadata based on keywords and time via MetadataSearchService API.
Database	Stores the metadata of the extracted, processed and analyzed data sets (Metadata store implementation in Fig. 2).
MetadataSearchAdapter, MetadataCollectionAdapter	Adapters for translation of interaction between entities of a big data system and database.
Metadata	Contains definitions of quality, administrative, descriptive, process, and navigational metadata for reception via MetadataCollectionService API and storage to database.
MetadataUIContext	Contains metadata definitions (including additional quality attribute parameters) for publishing via MetadataSearchService API.
MetadataValidator	Validates the metadata received via MetadataCollectionService.
MetadataQualityManagement	Manages quality metadata for the data sets utilizing the company's quality policies.
MetadataQualityPolicyManager	Manages the company's quality policies, containing both the organizational and decision making policies.
MetadataQualityEvaluationService	Responsible for evaluating the values for the metrics of the quality attributes.
MetadataUI	Provides the user interface for visualization of metadata to the end-user.
BigDataSystemComponent	Component of a big data system, which provides metadata to metadata management.

the analyzed data set. In this step, provided information includes navigational, process, descriptive and administrative metadata.

2.2 Metadata management notifies Quality management to create appropriate quality metadata for the analyzed data set.

2.3 Quality management notifies the Quality policy manager to select the appropriate metadata quality attributes for the source type 'social media data' according to the quality policy.

2.4 The Quality policy manager returns the appropriate quality attributes for the analyzed data set

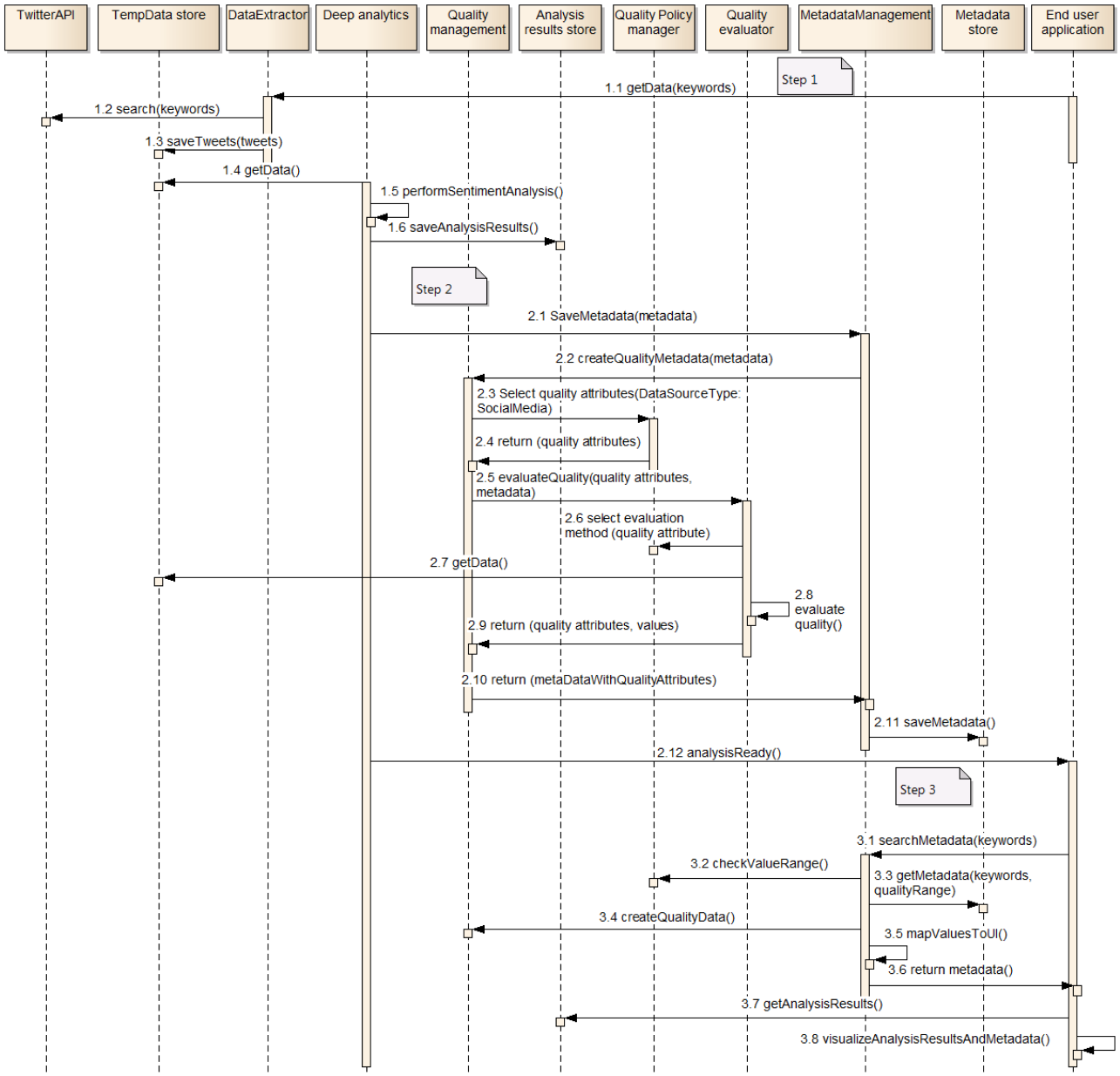

FIGURE 8. **Data and quality metadata management in big data architecture.**

(defined in organizational policy): timeliness and relevancy.

2.5 Quality management asks the Quality evaluator to evaluate the quality attributes.

2.6 For each quality attribute, the Quality evaluator checks for the appropriate metrics, evaluation methods and techniques defined in the organizational quality policy from the Quality policy manager.

2.7 The Quality evaluator fetches the data set (tweets) based on (navigational) metadata, which indicates location of the data set.

2.8 The Quality evaluator evaluates the following quality attributes: Timeliness is evaluated based on the timestamp of the metadata. Relevancy is determined based on the quality of the sentiment analysis algorithm (i.e. performance/quality of the analysis method), which is included into the metadata (in process metadata).

2.9 The Quality evaluator returns metadata with the quality attributes with values to Quality manager.

2.10 The Quality manager returns the values to Metadata management.

2.11 Metadata management writes the values into the quality metadata and saves the metadata into the Metadata store.

2.12 Deep analytics notifies the End-user application about a new analyzed data set.

Step 3 (Visualizing the Data to User for Decision Making): In this phase, the metadata is searched from the database for presentation to the end-user for decision making purposes. In this case example, the selected quality attributes include timeliness and relevancy. The relevant data is visualized to the end-user; the decision making policy defines the valuable data for the decision making by selecting only the data sets with the adequate quality attribute values. These policies are defined case-specific and applicable to the certain situation at hand. By changing the value range in a policy, the data sets with lower quality values can be selected. The following describes Step 3 of Fig. 8 in more detail:

3.1 The end-user (in this case; a decision maker) manages the data analysis through end-user application. An end-user wants to search interesting data sets with user-defined keywords (e.g. ''sentiment analysis'' and ''Product X''). The End-user application asks Metadata management to search the semantic keywords that are saved in the navigational metadata. The navigational metadata includes a list of semantic tags or keywords identifying the data set.

3.2 Metadata management checks the minimum values of the quality attributes of the data sets to be selected for the analysis (defined in decision making policy) from the Quality policy manager. For example, the selected data sets may not be older than one month, their relevancy must be at least 0.9 and believability must be at least 0.6. (value range $0 \ldots 1$).

3.3 Metadata management selects metadata sets, which include provided keywords and exceed the minimum values for the quality attributes from the Metadata store.

3.4 The timeliness attribute is recreated based on the current time.

3.5 Numerical values of quality attributes are mapped into human readable text for UI representation (e.g. timeliness value $> 0.7 \rightarrow$ 'very recent').

3.6 Metadata management returns the metadata to the end-user application.

3.7 The End-user application fetches the sentiment analysis data sets from the Analysis results store based on the (navigational) metadata.

3.8 Sentiment analysis results and metadata are visualized for the end-user. In the case example, the end-user prefers high relevancy of data prior to timeliness; thus the results are visualized in order of their relevancy.

As the data is visualized to the end-user, the end-user receives real time, validated information to support decision making. The company's decision makers then decide which actions to take. The company can have different levels of decision makers; the information is visualized according to the decision making policy. The decision making still requires a human and his/her expertise, and is assisted by the knowledge that the company has achieved (defined in the decision making policies).

In the case example, the data end-user receives the analysis results in order of their relevancy to the situation at hand. The user receives the positive and/or negative sentiment about the product, and uses this information, for example, to detect what kind of product features are desired and thus could be implemented and which features are negative and could be improved.

V. VALIDATION OF THE SOLUTION

The objective of the case example was to demonstrate the metadata and quality management with a social media use case. The implementation was conducted under DIGILE's Need for Speed (N4S) program[5] in collaboration with an industrial partner and VTT.[6] Metadata management was implemented and integrated with a big data use case as follows: The case company (company X) has built (into a public cloud) a system, which extracts tweets based on user-defined keywords, and performs sentiment analysis and visualization with a user interface (steps 1.1 - 1.6 in Fig. 8). We (VTT) provided the metadata management implementation, which is executed in VTT's separate, private cloud, and which provides a REST API for the big data system. The software implementation of the big data system was instrumented with calls to the metadata management interface (by company X) for transmitting of metadata information (step 2.1 in Fig. 8). VTT implemented the rest of the steps of Fig. 8 (from step 2.2. ahead), and also built a separate user interface into the private cloud for visualizing collected metadata for both organizations.

Currently, company X provides metadata information of extracted Twitter data sets, which is utilized as a basis for sentiment analysis. DataSourceType indicates the type of collected data sets, which can be utilized for determination of the relevancy attribute (step 2.8 in Fig. 8). Timeliness is determined based on the provided timestamp at the time of metadata extraction by comparison to the current time (step 3.4 in Fig. 8).

A. IMPLEMENTATION

When metadata management architecture was implemented, the technology choices, at least for metadata storage and API to the big data system, had to be determined (MetadataCollectionService and MetadataSearchService in Fig. 7). The technology choices are described and discussed in the following:

1) DATABASE FOR METADATA

Cassandra [46]. Metadata is saved into a column family, where a compound primary key for data was created based on a timestamp, and a parameter of descriptive metadata. An index had to be created into navigational metadata to

[5]http://www.n4s.fi/en/
[6]VTT Technical Research Centre of Finland.

enable searching based on keywords. Also, filtering has to be enabled in database queries based on keywords (with 'allow filtering'). This may lead to sub-optimal query latency, when compared to queries implemented with the primary key, which is very efficient in Cassandra [46]. Alternatively, a document oriented database (e.g. MongoDB) could be selected for storage of metadata due to the document structure of metadata.

2) METADATA API

XML over REST with Jersey. Alternatively, SOAP could have been selected as an implementation technology instead of REST. Earlier performance tests have indicated that REST has better performance than SOAP [47], [48]. The differences between REST and SOAP have been compared at the service level [48].

3) XML FORMAT VALIDATOR

Hibernate Validator. XML is an industry standard for platform-independent messaging. Alternatively, exchanged messages over REST could have been implemented with JSON. Differences between XML and JSON formats have been analyzed in terms of schema interoperability, serialization format, and message protocol [49].

B. VALIDATION

Initially, company X had an implementation of the social media use case. A requirement was to introduce only small changes to their existing software, which would enable extraction of metadata. Thus, VTT implemented metadata management architecture, which provided a REST interface for enabling straightforward instrumentation of software. One practical hindrance in the integration was the requirement for allowing cross-origin resource sharing [50]. This was caused by company X using a web browser within the enterprise domain, whereas data extraction and analysis was executed in the public cloud domain. In practice, the Access-Control-Allow-Origin header was needed in a HTTP response (to a HTTP OPTIONS request) for allowing access from the public cloud domain for extraction of metadata (a HTTP POST) with the web browser UI.

Metadata management implementation required about one month of development time, whereas instrumentation of a big data use case required one day of development time. No significant obstacles were discovered regarding the technological choices (see previous sub-section), when implementing extraction and search functionality of metadata. However, a more detailed analysis of performance and functionality may be needed, when the system is developed further with additional functionality.

The validation of the research solution was divided by company X focusing on big data use case R&D, while VTT designed and implemented metadata management architecture. Responsibilities were clearly divided in order to enable both organizations to focus on development of their software assets. REST API facilitated independent work on the activities by the organizations, and agreement of a common interface for integration. For the resource reasons, the existing demo of company X was used as a basis for implementation.

Currently, all steps of Fig. 8 have been implemented with the following limitations:

- Only one quality policy is implemented at this moment.
- The data in the case example was confidential data of the case company. This restricted the implementation of Step 2.7.
- Timeliness (time range) and keywords can be specified in the UI for searching of metadata (in step 3.1).

C. COMPARISON WITH RELATED WORKS

Only few works exists that relate to our solution. A quality evaluation framework for a big data pre-processing service is introduced in [51]. The framework is a generic solution that can be applied to different application domains, such as business, e-Health, IoT and social web. The quality evaluation pre-processing service is activated by a request with input data sources, output data destinations and a data quality profile. Each data input source has a data quality profile that contains reference to the actual data sources, output data and data quality rules. The pre-processing service includes the following architectural components: pre-processing activity selection, techniques selection, data quality selection, data profile optimization, data quality profile execution, quality control and data quality profile adapter. The quality evaluation service works iteratively by executing the defined processing activities and using the data profile adapter to change the data quality profile and notify the user about failed rules with suggestions on quality profile rules for better results. When compared to our solution, the main difference is the scope and focus. The scope of the proposed solution covers the latter part of the Data loading and pre-processing phase of our pipeline architecture introduced in Fig. 2. Our intent is to provide an architectural solution for managing quality of data in different phases of big data processing. Also, our solution focuses on using social media data in business decision making. Thus, all quality attributes of big data are not covered in our solution or in this quality framework.

Data quality centric big data architecture for federated sensor service clouds is introduced in [52]. The main contribution is the data quality (DQ)-aware virtualization of sensor services by enhancing each sensor feed's metadata with data quality attributes. The main components of the architecture are the DQ services catalog and DQ monitoring and adaptation component. Analysis is made in two phases: online feed analysis and batch analysis. The data quality model includes the following attributes: accuracy, error rate, availability, timeliness and validity. The main differences are in the architecture style and data quality model. This architecture focuses on connecting physical data sources to applications by applying a domain-specific data quality model. On the contrary, our solution focuses on big data processing and intends to manage the quality of unstructured social media data in each processing phase and applying quality policies

for adapting a quality model to the evaluation phase and data user's situation.

D. FUTURE DEVELOPMENT DIRECTIONS

The following development targets have been identified:

1. Implementation of several quality policies: Currently, each organizational policy is associated with the DataSourceType and one or more quality attributes. The QualityPolicyManager is responsible for initialization of the organizational quality policies. In the future more organizational quality policies could be defined for different social media types.
2. Evaluation of several quality attributes: Currently, our work is mainly an architecture for creation of quality aspects as part of overall metadata in a big data system (in the context of social media). Initially, the timeliness attribute provided a value based on the timestamp. In the future, algorithms will be developed, implemented, and validated for determination of several quality attributes in order to improve the utility of the solution.
3. Data/information search and user interface to quality management: The quality policies must be visualized to the user; the user must be able to, for example, update the quality policies, change the rules or add new acceptable data sources. The search based on other quality attributes must be implemented as well.

First of all, the different types of social media data (e.g. data from Twitter, Facebook or Instagram) should be able to be used together. Therefore, the definition of quality metrics for different types of social media data and rules for how to apply the properties of data quality metrics must be rationalized. Finally, the solution must be applied to different application domains and with different decision support systems to see how the quality attributes and rules are managed in different cases.

VI. CONCLUSIONS

This paper introduced a solution to evaluate the quality of data for business decision making purposes. The quality of data is evaluated in each data processing phase of the big data architecture with the help of quality metadata and quality policies. The solution may be adapted to different contexts, enabling the user to select the applicable quality attributes, evaluate them and apply them in a suitable way into a certain situation. The solution is also extendable; it allows inserting new data sources and data sets for data extraction, as well as new metrics and algorithms for data evaluation. The metadata enables location, retrieval and management of all the data sets, and the quality attributes and their values in metadata enable detection of the quality and value of data in a certain situation.

The solution was demonstrated with a case example where a company finds out the level of customer satisfaction regarding the quality of a product utilizing social media data. The solution was implemented with an industrial partner using a standard interface, which facilitated independent work of the company and the research organization, and functioned as a good communication tool for agreement with the integration. Several development targets were identified when demonstrating the solution. First of all, support for automating the quality attribute evaluation is required. The (semi-) automated adaptation of the organizational and decision making policies is required as well. However, the more knowledge the company achieves, the more the decision making process can be automatized with the help of quality policies.

At this moment the quality evaluation is limited to only a few quality attributes; the purpose is to extend the quality evaluation to include more quality attributes. One of the most important development targets is, however, to include other data source types, such as customer feedback data, product data and market analysis, to the quality evaluation to achieve 'customer insight' into business decision making.

REFERENCES

[1] S. R. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a Web of open data," in *The Semantic Web* (Lecture Notes in Computer Science), vol. 4825. Berlin, Germany: Springer-Verlag, 2007, pp. 722–735.

[2] A. Immonen, M. Palviainen, and E. Ovaska, "Requirements of an open data based business ecosystem," *IEEE Access*, vol. 2, pp. 88–103, Feb. 2014. DOI: 10.1109/ACCESS.2014.2302872

[3] S. Bhatia, J. Li, W. Peng, and T. Sun, "Monitoring and analyzing customer feedback through social media platforms for identifying and remedying customer problems," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining (ASONAM)*, Aug. 2013, pp. 1147–1154.

[4] F. Antunes and J. P. Costa, "Integrating decision support and social networks," *Adv. Human-Comput. Interact.*, vol. 2012, Jan. 2012, Art. ID 9.

[5] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer feedback and data collection techniques in software R&D: A literature review," in *Software Business* (Lecture Notes in Business Information Processing), vol. 210. Berlin, Germany: Springer-Verlag, 2015, pp. 139–153.

[6] R. Ferrando-Llopis, D. Lopez-Berzosa, and C. Mulligan, "Advancing value creation and value capture in data-intensive contexts," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 5–9.

[7] P. Pääkkönen and D. Pakkala, "Reference architecture and classification of technologies, products and services for big data systems," *Big Data Res.*, Jan. 2015. DOI: 10.1016/j.bdr.2015.01.001

[8] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and framework for data and information quality research," *J. Data Inf. Quality*, vol. 1, no. 1, 2009, Art. ID 2.

[9] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[10] J. R. C. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts, "Information quality and trustworthiness: A topical state-of-the-art review," in *Proc. Int. Conf. Comput. Appl. Netw. Secur. (ICCANS)*, Malé, Maldives, 2011, pp. 492–500.

[11] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, Jan. 2015.

[12] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, no. 2, pp. 1–10, 2015.

[13] B. Ramesh, "Big data architecture," in *Studies in Big Data*, vol. 11, H. Mohanty *et al.*, Eds. India: Springer-Verlag, 2015. DOI: 10.1007/978-81-322-2494-5_2

[14] M. Chen *et al.*, "Data, information, and knowledge in visualization," *IEEE Comput. Graph. Appl.*, vol. 29, no. 1, pp. 12–19, Jan./Feb. 2009.

[15] *Software Engineering—Product Quality. Part 1: Quality Model*, ISO/IEC Standard 9126-1, ISO/IEC, 2001, p. 25.

[16] National Information Standards Organization, *Understanding Metadata*. Bethesda, MD, USA: NISO Press, 2004.

[17] W3C. (2007). *Web Services Policy 1.5—Framework (W3C Recommendation)*. [Online]. Available: http://www.w3.org/TR/ws-policy/
[18] *Quality Management Systems—Requirements*, ISO Standard 9001:2008, ISO, 2008.
[19] I. Gorton and J. Klein, "Distribution, data, deployment: Software architecture convergence in big data systems," *IEEE Softw.*, vol. 32, no. 3, pp. 78–85, May/Jun. 2015.
[20] N. Foshay, A. Mukherjee, and A. Taylor, "Does data warehouse end-user metadata add value?" *Commun. ACM*, vol. 50, no. 11, pp. 70–77, 2007.
[21] B. E. Bargmeyer and D. W. Gillman, "Metadata standards and metadata registries: An overview," in *Proc. Int. Conf. Establishment Surv. II*, Buffalo, NY, USA, 2000, pp. 1–10.
[22] *Common Warehouse Metamodel (CWM) Specification*, OMG document ad/99-09-01, OMG, 1999. [Online]. Available: http://www.omg.org
[23] *Information Technology—Metadata Registries (MDR)—Part 3: Registry Metamodel and Basic Attributes*, ISO/IEC Standard 11179-3:2003(E), International Organization for Standardization, Geneva, Switzerland, 2003.
[24] National Information Standards Organization, *The Dublin Core Metadata Element Set, Version 1.1*. Bethesda, MD, USA: NISO Press, 2012. [Online]. Available: http://www.dublincore.org/documents/dces
[25] *Software Engineering—Product Quality. Part 2: External Metrics*, ISO/IEC Standard TR 9126-2:2003, 2003.
[26] *Software Engineering—Product Quality. Part 3: Internal Metrics*, ISO/IEC Standard TR 9126-3:2003, 2003.
[27] A. Immonen and E. Niemelä, "Survey of reliability and availability prediction methods from the viewpoint of software architecture," *Softw. Syst. Model.*, vol. 7, no. 1, pp. 49–65, 2008.
[28] E. Ovaska, A. Evesti, K. Henttonen, M. Palviainen, and P. Aho, "Knowledge based quality-driven architecture design and evaluation," *Inf. Softw. Technol.*, vol. 52, no. 6, pp. 577–601, 2010.
[29] E. Niemelä and A. Immonen, "Capturing quality requirements of product family architecture," *Inf. Softw. Technol.*, vol. 49, nos. 11–12, pp. 1107–1120, 2007.
[30] R. Kazman, M. Klein, and P. Clements, "ATAM: Method for architecture evaluation," Carnegie Mellon Univ., Softw. Eng. Inst., Pittsburgh, PA, USA, Tech. Rep. CMU/SEI-2000-TR-004, Aug. 2000. [Online]. Available: http://resources.sei.cmu.edu/asset_files/TechnicalReport/2000_005_001_13706.pdf
[31] L. Dobrica and E. Niemelä, "A survey on software architecture analysis methods," *IEEE Trans. Softw. Eng.*, vol. 28, no. 7, pp. 638–653, Jul. 2002.
[32] Y. Gil and D. Artz, "Towards content trust of Web resources," *Web Semantics, Sci., Services, Agents World Wide Web*, vol. 5, no. 4, pp. 227–239, 2007.
[33] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in *Proc. 5th VLDB Workshop Secure Data Manage.*, vol. 5159. 2008, pp. 82–98.
[34] F. Naumann and C. Rolker, "Assessment methods for information quality criteria," in *Proc. 5th Int. Conf. Inf. Quality*, Boston, MA, USA, 2000, pp. 148–162.
[35] J. R. C. Nurse, I. Agrafiotis, S. Creese, M. Goldsmith, and K. Lamberts, "Building confidence in information-trustworthiness metrics for decision support," in *Proc. 12th IEEE Int. Conf. Trust, Secur., Privacy Comput. Commun. (TrustCom)*, Melbourne, VIC, Australia, Jul. 2013, pp. 535–543.
[36] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.
[37] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. Int. Conf. Web Search Data Mining (WSDM)*, 2008, pp. 183–194.
[38] C. Bizer, "Quality-driven information filtering in the context of Web-based information systems," Ph.D. dissertation, Dept. Econom. Sci., Freie Univ. Berlin, Berlin, Germany, 2007.
[39] C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," *Web Semantics, Sci., Services, Agents World Wide Web*, vol. 7, no. 1, pp. 1–10, 2009.
[40] S. S. Rahman, S. Creese, and M. Goldsmith, "Accepting information with a pinch of salt: Handling untrusted information sources," in *Security and Trust Management* (Lecture Notes in Computer Science), vol. 7170. Berlin, Germany: Springer-Verlag, 2011, pp. 223–238.
[41] E. Bertino and H.-S. Lim, "Assuring data trustworthiness—Concepts and research challenges," in *Secure Data Management* (Lecture Notes in Computer Science), vol. 6358. Berlin, Germany: Springer-Verlag, 2010, pp. 1–12.
[42] E. Niemelä, A. Evesti, and P. Savolainen, "Modeling quality attribute variability," in *Proc. 3rd Int. Conf. Eval. Novel Approaches Softw. Eng.*, Funchal, Portugal, 2008, pp. 169–176.
[43] V. Luukkala and I. Niemelä, "Enhancing a smart space with answer set programming," in *Semantic Web Rules*, M. Dean, J. Hall, A. Rotolo, and S. Tabet, Eds. Berlin, Germany: Springer-Verlag, 2010, pp. 89–103.
[44] W3C. (2012). *SPARQL Query Language for RDF*. W3C Recommendation. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query
[45] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.
[46] DataStax. *CQL for Cassandra 2.2*. [Online]. Available: http://docs.datastax.com/en/cql/3.3/cql/cqlIntro.html, accessed Aug. 10, 2015.
[47] T. Aihkisalo and T. Paaso, "Latencies of service invocation and processing of the REST and SOAP Web service interfaces," in *Proc. IEEE 8th World Congr. Services*, Jun. 2012, pp. 100–107.
[48] G. Mulligan and D. Gracanin, "A comparison of SOAP and REST implementations of a service based interaction independence middleware framework," in *Proc. Winter Simulation Conf.*, Austin, TX, USA, Dec. 2009, pp. 1423–1431.
[49] J. Delgado, "Service interoperability in the Internet of Things," in *Internet of Things and Inter-Cooperative Computational Technologies for Collective Intelligence* (Studies in Computational Intelligence), vol. 460. Berlin, Germany: Springer-Verlag, 2013, pp. 51–87.
[50] A. van Kesteren. (2014). *Cross-Origin Resource Sharing*. W3C Recommendation. [Online]. Available: http://Www.W3.Org/TR/Access-Control/
[51] I. Taleb, R. Dssouli, and M. A. Serhani, "Big data pre-processing: A quality framework," in *Proc. IEEE Int. Congr. Big Data*, New York, NY, USA, Jun./Jul. 2015, pp. 191–198.
[52] L. Ramaswamy, V. Lawson, and S. V. Gogineni, "Towards a quality-centric big data architecture for federated sensor services," in *Proc. IEEE Int. Congr. Big Data*, Santa Clara, CA, USA, Jun./Jul. 2013, pp. 86–93.

ANNE IMMONEN received the M.Sc. degree in information processing science from the University of Oulu, Finland, in 2002. She is currently a Research Scientist with the VTT Technical Research Centre of Finland. Her main research interests include reliability in service engineering, in particular, in the context of digital service ecosystems. Her current research interests include the data and service ecosystems, big data, and the quality and trustworthiness of data.

PEKKA PÄÄKKÖNEN received the M.Sc. degree in information technology from the University of Oulu, Finland, in 2002. He is currently a Senior Research Scientist with the VTT Technical Research Centre of Finland. His research interests include distributed computing, big data technologies, databases, and software performance.

EILA OVASKA received the Ph.D. degree from the University of Oulu, in 2000. Prior to 2000, she was a Software Engineer, a Senior Research Scientist, and the Leader with the Software Architectures Group, VTT Technical Research Centre of Finland. Since 2001, she has been a Research Professor with the VTT Technical Research Centre of Finland and an Adjunct Professor with the University of Oulu. She has co-authored over 150 scientific publications. Her current areas of interest are service architectures, self-management systems, and knowledge oriented service engineering. She has acted as a Workshop and Conference Organizer and Reviewer for scientific journals and conferences.