Big Data Research ••• (••••) •••–•••

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems

Pekka Pääkkönen [1], Daniel Pakkala [1]

VTT Technical Research Centre of Finland, Kaitoväylä 1, 90570, Oulu, Finland

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 18 December 2014
Accepted 11 January 2015
Available online xxxx

Keywords:
Big data
Reference architecture
Classification
Literature survey

ABSTRACT

Many business cases exploiting big data have been realised in recent years; Twitter, LinkedIn, and Facebook are examples of companies in the social networking domain. Other big data use cases have focused on capturing of value from streaming of movies (Netflix), monitoring of network traffic, or improvement of processes in the manufacturing industry. Also, implementation architectures of the use cases have been published. However, conceptual work integrating the approaches into one coherent reference architecture has been limited. The contribution of this paper is technology independent reference architecture for big data systems, which is based on analysis of published implementation architectures of big data use cases. An additional contribution is classification of related implementation technologies and products/services, which is based on analysis of the published use cases and survey of related work. The reference architecture and associated classification are aimed for facilitating architecture design and selection of technologies or commercial solutions, when constructing big data systems.

1. Introduction

Many big data use cases have been realised, which create additional value for companies, end users and third parties. Currently, real time data is gathered from millions of end users via popular social networking services. For example, LinkedIn [1] collects data from users, and offers services such as "People you may know", skill endorsements or news feed updates to end users based on analysis of the data. Another example is Netflix, which uses big data for providing recommendations and ranking related services to customers [2]. Twitter uses collected data for real time query suggestion and spelling corrections of their search algorithm [3]. Analysis of collected data also increases understanding of consumers, which is an important asset for the big data companies. Value from data can also be extracted with other applications such as monitoring of network traffic [4] or improving manufacturing process of digital displays [5].

A wide variety of technologies and heterogeneous architectures have been applied in the implementation of the big data use cases. The publications have mainly concentrated on describing architectures of individual contributions by large big data companies such as Facebook [6] or LinkedIn [1]. On the other hand, architectural work combining the individual reports into one coherent reference architecture has been limited, although the first contributions have been made [7–10]. Technology independent reference architecture and categorization of related implementation technologies and services would be valuable for research and development of big data systems.

The contribution of this paper is reference architecture for big data systems, and classification of related technologies and products/services. First, big data research, reference architectures, and use cases are surveyed from literature. Subsequently, the design of reference architecture for big data systems is presented, which has been constructed inductively based on analysis of the presented use cases. Finally, a classification is provided for the purpose of creating an overall picture of big data research, related technologies, products, and services.

The structure of the paper is as follows: Material and methods of the study are described in Section 2. Theoretical background is provided in Section 3. Design and construction of the reference architecture is presented in Section 4. Classification of big data technologies and commercial products/services, and survey of related work are provided in Section 5. The results are analysed in Section 6 and discussed in Section 7. A conclusion is provided in Section 8. The appendices include: a detailed description of the reference architecture (Appendix A), a detailed description of the

E-mail addresses: pekka.paakkonen@vtt.fi (P. Pääkkönen), daniel.pakkala@vtt.fi (D. Pakkala).
[1] Tel.: +358 207222299.

http://dx.doi.org/10.1016/j.bdr.2015.01.001
2214-5796/

research method (Appendix B), and references to surveyed commercial products and services (Appendix C).

2. Material and methods

The overall goal of this work is to facilitate realisation of big data systems. When a big data system is realised, important considerations include architecture design of the system, and utilization of underlying technologies and products/services [11]. The goals of this work are: a.) design technology independent reference architecture for big data systems b.) classify related technologies and products/services with respect to the reference architecture.

Reference architecture would be useful in the following ways: It should facilitate creation of concrete architectures [12], and increase understanding as an overall picture by containing typical functionality and data flows in a big data system. Classification of technologies and products/services should facilitate decision making regarding realisation of system functionalities. Also, it would be important to understand architecture and performance characteristics of related technologies. The following research questions are posed:

The first research question: *What elements comprise reference architecture for big data systems?*
The second research question: *How to classify technologies and products/services of big data systems?*

The reference architecture for big data systems was designed with inductive reasoning based on the published use cases described in Section 3.3 (research question 1). Particularly, functionality, data flows, and data stores of implementation architectures in seven big data use cases were analysed. Subsequently, reference architecture was constructed based on the analysis. The method for reference architecture design is described in detail in the Appendix. A literature survey was used for answering to the second research question.

3. Theory

Section 3.1 presents earlier surveys of big data. Research on big data reference architectures is presented in Section 3.2. The latest reports of big data use cases are introduced in Section 3.3. Finally, a summary of related work in presented in Section 3.4.

3.1. Big data research

Begoli [13] conducted a short survey of state-of-the-art in architectures and platforms for large scale data analysis. The survey covered adoption of related technologies, platforms for knowledge discovery, and architectural taxonomies. Chen et al. presented a comprehensive survey of big data [14]. The topics of the survey covers related technologies, generation and acquisition of data, storage, applications, and outlook to the future. Chen and Zhang also surveyed big data [11]. Their work focused on big data opportunities and challenges, techniques and technologies, design principles, and future research. Wu et al. provided a framework for big data mining [15]. The authors proposed HACE (Heterogeneous, Autonomous sources, Complex and Evolving relationships among data) theorem for characterizing big data. The authors also presented a three layer framework for big data processing, which is comprised of big data mining platform, semantics and application knowledge, and mining algorithms. Finally, Cuzzocrea et al. discussed Online Analytical Processing (OLAP) over big data, big data posting and privacy as part of big data research agenda [16].

3.2. Reference architecture for big data systems

A framework for design and analysis of software reference architectures has been presented [12]. The framework contains a multi-dimensional classification space, and five types of reference architectures. It is claimed that architecture design based on the classified reference architectures should lead to better success. Also, empirically-grounded design of software reference architectures has been presented [17]. The design approach is based on expected empirical material gathered with interviews, questionnaires, and document analysis. The procedure is a step-wise process, which consists of deciding a type for the reference architecture, selection of design strategy, empirical acquisition of data, construction of reference architecture, enabling of variability, and evaluation (see Appendix B for details).

Service-oriented reference architecture has been defined for enterprise domain [18]. However, in the big data context, there exist only few architecture proposals. Schmidt and Möhring [8] presented a service and deployment model for implementing big data pipeline to the cloud domain. Demchenko et al. presented a Big Data Architecture Framework, which consists of high-level description of big data lifecycle and infrastructure [9]. Doshi et al. presented reference architectures for integration of SQL and NewSQL databases in order to support different growth patterns in enterprise data traffic [19]. Zhong et al. proposed and validated big data architecture with high-speed updates and queries [20]. The architecture consists of in-memory storage system and distributed execution of analysis tasks. Cuesta proposed tiered architecture (SOLID) for separating big data management from data generation and semantic consumption [10]. Generalized software architecture was proposed for predictive analytics of historical and real-time temporally structured data [89]. Meier conducted design of reference architecture covering functionality in realised big data use cases (Master's Thesis [7]). The author initially defined requirements for reference architecture, conducted architecture design, and validated the presented architecture against published implementation architectures of Facebook, LinkedIn, and Oracle. The design was conducted in the empirically-grounded design framework for reference architectures [17,12].

3.3. Big data use cases

Many big data use cases have been published. Facebook, Twitter, and LinkedIn are examples in the social network application domain. Facebook collects structured and stream-based data from users, which is applied for batch-based data analysis [6]. Data scientists at Facebook can specify ad hoc analysis tasks in production or development environments for getting deep insight to the data. LinkedIn [1] also collects structured and stream-based data, which is analysed in development and production environments. Additionally, LinkedIn provides new services (e.g. "People you may know") for end users based on data analysis [1]. Twitter [3,21,22] handles mainly tweets, which have real-time processing requirements. Twitter also provides new services for end users e.g. "Who to follow" [23].

Netflix is a commercial video-streaming service for end users. Netflix collects user events, which are processed and analysed in online, offline, and nearline environments [2]. Video recommendations are provided for end users based on real time data analysis.

Also, network traffic has been analysed for getting value from data. BlockMon [4,24] is a high performance streaming analytics platform, which has been used for telemarketer call detection based on Call Data Records (CDR). Another application is monitoring of network traffic for execution of ad hoc Hadoop/MapReduce tasks [25,26]. The primary applications are web traffic analysis and

Please cite this article in press as: P. Pääkkönen, D. Pakkala, Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.01.001