Towards an Architecture for Big Data-Driven Knowledge Management Systems

Full papers

Thang Le Dinh
Université du Québec à Trois-Rivières
Thang.Ledinh@uqtr.ca

Thuong-Cang Phan
Université du Québec à Trois-Rivières
Thuong-Cang.Phan@uqtr.ca

Trung Bui
Adobe research
Bui@adobe.com

Abstract

Nowadays, knowledge management systems are confronted with a variety and unprecedented amount of data, resulting from big data sources. A new generation of knowledge management systems for exploring and exploiting big data becomes a major need for organizations. For this reason, the paper proposes a novel service-oriented architecture for big data-driven knowledge management systems. The purpose of this research is to support organizations to leverage their knowledge-based assets for improving decision-making and facilitating organizational learning. The proposed architecture is based on the principles of design science research, including a set of constructs, a model and a method. The design evaluation is presented based on the analytical evaluation method. By applying the architecture, an organization can manage and govern business and digital transformation, setting them apart from their competitors.

Keywords

Big Data, Big Data Analytics, Knowledge Management System, Service Orientation.

Introduction

The development and use of Knowledge Management Systems (KMSs) are currently having a direct and dramatic impact on business decisions and processes in modern and networked organizations (Kapuruge 2011). KMSs render organizations more competitive to grasp more business opportunities (Alavi and Leidner 2001; Djordjevic-Boljanovic et al. 2013). However, these KMSs are currently confronted with a variety and unprecedented amount of data, resulting from different business and IT-based services, called “big data” (Chen et al. 2012). Big data provides high-volume, high-velocity and high-variety information assets that leads to a revolution of transforming traditional organizations into knowledge-intensive ones, called Data-Driven Organizations (DDOs). Consequently, knowledge discovered in DDOs needs to be translated from big data into organizational knowledge to aid managers in making decision and in improving performance (Chen et al. 2012). Despite the fact that big data research has recently gained rapid growth, there is a lack of frameworks and architectures that enable DDOs to capture the value of big data in a systematic manner, especially for promoting organisational learning (Wang et al 2015; Olivo et al. 2016). Indeed, one of the most important challenges for KMSs today is to be able to deal with big datasets that are required to be updated frequently or continuously. Therefore, a new generation of KMSs that is able to handle efficiently big data sources becomes an essential tool for organizations, especially DDOs.

In fact, most of the recent studies, which are related to the integration of big data and KMSs, have separately concentrated on specific aspects of knowledge management such as business intelligence and business analytics (Chen et al. 2012), data mining and knowledge discovery (Begoli and Horey 2012; Wu et al. 2014). These studies have strongly focused on knowledge exploration, but have not been fully supported knowledge exploitation yet. There has been little attempt to take into account the impact of big data on the whole process of organizational knowledge management and the trend of service orientation.
As a result, this research aims at proposing a novel service-oriented architecture for big data-driven knowledge management systems, hereafter called BDD-KMS architecture. This rest of the paper is structured as follows. Section 2 provides the summary of relevant literature. Section 3 describes the essentials of the research design and positions our paper with respect to the related work. Section 4 introduces the proposed architecture according to the design science research, including a set of constructs, a model and a method. Section 5 demonstrates a design evaluation of the architecture. Section 6 outlines the conclusions and some ideas for further research.

**Literature Review**

**Knowledge Management Systems and Big Data**

In this research context, *data* consists of traditional data and big data. *Knowledge* is constituted from *knowledge objects*, which are classified on the basis of their level of development, that is, as data, information, knowledge or wisdom (Bierly III et al. 2000). *Knowledge management* (KM) is defined as organizational activities related to knowledge artefacts in which a learning process has occurred, and intellectual capital is accumulated and developed. *Knowledge artefacts* are data or information collected from existing systems, social networks, telecommunication networks, and mobile services. *Knowledge management systems* represent a specific type of information systems applied to handle organizational knowledge (Alavi and Leidner 2001), that includes activities such as knowledge capture, knowledge organization, knowledge transfer and knowledge application (Le Dinh et al. 2015). The backbone of KMSs is the *knowledge architecture*, which is the application of information architecture to knowledge management that supports and enhances the KM activities.

Big data brings a great opportunity but also a big challenge for implementing KMSs in DDOs (Beyer and Laney 2012). The importance of big data does not revolve around how much data organizations have today, but what business value they can distill from this data in the best manner and within the shortest response time. In the global competitive environment, organizations, which are able to leverage effectively big data through KMSs, can differentiate themselves from their rivals. However, there is a great challenge for DDOs for handling the immense volumes of data, which are being continuously generated on an hourly basis. Hence, a huge computing power for analytics, which is required to handle the unprecedented input, creates significant barriers for organizations to harness effectively the business value of big data.

**Service Orientation**

*Service Orientation* (SO) is a design paradigm for software applications constituted by services (Erl 2005). It offers critical features such as standardization, platform independence, well-defined interfaces, and various tools that support the integration of legacy and new systems. Accordingly, SO rapidly becomes a popular architectural approach used to implement distributed and loosely coupled systems (Yang 2013). *Service-oriented architecture* (SOA) is a design architecture for constructing information systems by the combination of services. The heart of SOA is a set of standard approaches for designing and sharing reusable services and an implementation of distributed and loosely coupled services. The SOA approach satisfies several business goals such as interoperability, maintainability, modifiability, extensibility, availability, reliability, security, and performance. These are the primary drivers for adopting SOA as a set of design principles for service-oriented systems.

**Related Work**

Conventional knowledge management systems were not prepared well for storing, processing and analyzing big data because they mainly use traditional databases (Jurisica et al. 2004). A survey of current research projects related to big data indicated that most of them have had only focused on specific aspects of KM such as big data management, business analytics, business intelligence, and decision-making (Chen et al. 2012). Recently, some studies related to big data and knowledge management have additionally mentioned on other aspects such as methods for building knowledge bases (Suchanek and Weikum 2014), data mining and knowledge discovery (Begoli and Horey 2012; Wu et al. 2014), knowledge fusion based on machine learning (Dong et al. 2014), deep learning (Yu 2013) and random walk inference (Lao et al. 2011), and real-time stream data processing (Esposito et al. 2015). However, these studies did not provide
a comprehensive architecture for supporting the complete process of knowledge development as well as the collaboration of their systems to support organizational learning (Wang et al. 2015; Olivo et al. 2016).

On the other hand, the service orientation approach was regarded as a flexible, agile, and powerful approach for implementing KMSs (Šaša and Krisper 2010; Dai and Rubin 2012). Service-oriented KMSs explored the management of resources as services, and exploited the flexibility of information systems. Unfortunately, these systems did not consider the opportunities being presented by big data. In addition, there is still a little concern for the knowledge fusion of traditional data and big data, as well as for the comprehensive process of knowledge development.

In order to overcome the above challenge, we propose a general architecture for BDD-KMSs based on service-oriented principles. Our architecture supports the whole knowledge management process in a DDO in order to provide a unified way of working, learning and innovating in the big data era.

**Research Design**

**Research Question**

The research question of this study is “How to implement effective KMSs in the big data era?” Consequently, we address two main objectives: i) Capturing and unifying the knowledge discovered from diverse big data and traditional data sources, and ii) Supporting the process of organizational knowledge management based on the service orientation perspective.

**Research Method**

This research proposes a novel service-oriented architecture, which can be considered as a design artifact. Accordingly, the design science research has been chosen to carry out this research that aims at developing a technology-based solution to overcome a business challenge (Hevner et al. 2004). Key elements of our research method are described as follows:

- Concerning the problem relevance, the business challenge addressed by this research is how to implement effective KMSs in the big data era, which facilitates the whole process of organizational knowledge management, including knowledge exploration and knowledge exploitation.

- Concerning the research rigor, the presented study has the theoretical foundations in knowledge management theory as well as in information system theory. In order to support the whole process of organizational knowledge development, we adopt a hybrid KM model that supports the both views of knowledge management: epistemology-oriented and ontology-oriented KM models (Gebert et al. 2003). The first views a knowledge object as an entity that can be deconstructed into discrete and relevant attributes; meanwhile, the second defines knowledge solely through their relationships with a constructed universe of discourse. Furthermore, the proposed architecture is constructed based on diverse systems and services at different layers of knowledge objects such as data, information, knowledge and understanding layers.

- Concerning the search process, we focus on the exploration, implementation, experimentation and iteration. We study the related work and trends related to big data and KMSs, then implement or experiment (or both) the different aspects of the architecture to determine the suitability for real-world applications. The cycle, which generates design alternatives and tests alternatives against business requirements, has been repeated several times during the period of 18 months by the research team and their business partners.

- Concerning the design artefact, the proposed architecture includes a set of constructs, a model and a method (Hevner et al. 2004). The constructs and model support the first objective of this research; meanwhile, the method supports the second one. Constructs are different types of concepts related to knowledge objects and their knowledge components in a data-driven organization and its business environment. A model is a set of statements expressing the relationships between constructs at different levels in the knowledge structure. A method is a set of activities that supports the process of organizational knowledge development.

- Concerning the design evaluation, the analytical evaluation method is used to validate the design artefact (Hevner et al. 2004). In particular, static analysis is used to examine the structure of the
proposed architect, and the architecture analysis is used to study the fit of the design artefact into current information system architecture.

**Service-Oriented Architecture for BDD-KMSs**

*Constructs of the BDD-KMS Architecture*

As mentioned above, *constructs* are different types of concepts related to knowledge objects and their knowledge components. As presented in Figure 1, “a knowledge object (KO) is a highly structured interrelated set of data or content, information, knowledge, and wisdom concerning some business situation”, which provides a viable approach for dealing with the situation (Bellenger 2004). In this study, the term “understanding” is used instead of the term “wisdom” since the research just focuses on the first level of understanding in which a DDO understands how to increase business values by using their knowledge and knowing. In other words, the proposed architecture is still at a level of knowledge-based decision support instead of the best decision-making support.

![Figure 1. Knowledge Development Levels in KMSs](image-url)

According to the structure view, a knowledge management process covers both knowledge exploration and knowledge exploitation. The process of knowledge exploration concerns the *capture* and *organization* of different knowledge (tacit and explicit) in the organizational memory, whereas the process of knowledge exploitation concerns the *transfer* and *application* of classified and organized explicit knowledge in the knowledge repository (Le Dinh et al. 2015). A knowledge object is considered as a collection of knowing, called knowledge components (Le Dinh et al. 2015). Knowledge components can be stratified in different levels ranging from lowest to highest: cognitive (know-what), conditional (know-when), situational (know-where), applied (know-how) and rule-of-thumb (know-why). A knowledge component can be explicit or tacit knowledge as well as individual or collective knowledge. Knowledge components of knowledge objects can be used and shared by using different knowledge conversion processes such as socialization, externalization, combination and internalization (Nonaka et al. 1996).

According to the behavior view, a life cycle of a knowledge object includes the following dynamic states: *Captured knowledge*, *Organized knowledge*, *Semantical knowledge* and *Situational knowledge*. Firstly, *data* is raw and simply exists in any form (Bellinger et al. 2004). A KO is at the *Captured knowledge* state if its data is captured and stored in the knowledge repository. Secondly, *information* is data that has been given meaning by way of relational connection (Bellinger et al. 2004). A KO is at the *Organized knowledge* state if its data is organized according to its knowledge components corresponding to the structure of knowledge (Le Dinh at al. 2014) to become useful information. A KO at this level can answer the simple questions such as “What”, “Who”, “Where”, “When”. Thirdly, *knowledge* is the appropriate application of information to facilitate organizational activities (Bellinger et al. 2004). A KO is at the *Semantical knowledge* state if its knowledge components corresponding to the structure of knowledge are linked to the knowledge components corresponding to the transition of knowledge (Le Dinh at al. 2014). A
KO at this level is able to provide an answer or guidelines for a “How” question. Fourthly, understanding is the process of by which a person can take knowledge and synthesize new knowledge from the previous held knowledge to make business decisions (Bellinger et al. 2004). A KO is at the Situational knowledge state if its knowledge components corresponding to the structure and transition of knowledge are linked to the knowledge components corresponding to the coherence of knowledge (Le Dinh at al. 2014) so that it can provide guidelines for a person to take necessary actions faced a specific business situation.

**Model of the BDD-KMS Architecture**

A model is defined as the statements expressing the relationships between knowledge components of knowledge objects. Knowledge components can be classified based on the key aspects of knowledge that forms the knowledge structure of a DDO: structure, transition, and coherence of knowledge (Le Dinh at al. 2014). The **structure of knowledge**, which is represented by the “know-what” knowledge component and its associated knowledge-components (know-who, know-when, know-where), describes knowledge that relates to a phenomenon of interest (Garud, 1997). Know-what in a DDO can relate to products, services, or markets. The **transition of knowledge**, which is represented by the “know-how” knowledge component, describes the understanding of the generative processes that constitute phenomena (Garud, 1997). Know-how in a DDO concerns the organizational processes, which are a feedback of the organization to the occurrence of an event or a situation. The **coherence of knowledge**, which is represented by the “know-why” knowledge component, that describes the understanding of the principles underlying phenomena. Know-why in a DDO concerns the business rules and guidelines, which help to make business decisions to face certain situations.

As presented in Figure 2, the service-oriented architecture consists of four layers: Data-as-a-Service, Information-as-a-Service (corresponding to the structure of knowledge), Knowledge-as-a-Service (corresponding to the transition of knowledge), and Business Process-as-a-Service layers (corresponding to the coherence of knowledge).

**Figure 2. A Service-Oriented Architecture for Big Data-Driven KMSs**

*Data-as-a-Service layer* (DaaS) includes data loading and ingestion components, a data repository, and data services. DaaS is a data collection, provision and distribution model in which big data is made available to higher-level layers and users over the network.

*Information-as-a-Service layer* (IaaS) contains batch and real-time processing components, a knowledge repository, and information services. IaaS structures and analyzes the disorder, disparate and obscure data from the batch and real-time data into linked, integrated, and semantic information that can be delivered on demand as a service. This information can be organized by tagging based on the corresponding knowledge components of knowledge objects. In addition, data files can be categorized and associated automatically or semi-automatically to a set of related concepts of the knowledge structure.
Knowledge-as-a-Service layer (KaaS) typically provides knowledge services supporting organizational activities such as knowledge discovery, collaboration, learning, decision-making, and control services (Jafari et al. 2009; Alavi and Leidner 2001). KaaS transfers the information from the IaaS layer into knowledge based on their semantic contexts and organizes this knowledge corresponding to the “know-how” knowledge component. Besides, it also provides the ability to analyze, learn, infer, and make the knowledge and experience available to the higher business layer and users.

Business Process-as-a-Service layer (BPaaS) includes process representation, which provides building blocks for aggregating the loosely-coupled services of the lower layers as a collaborative process aligned with business goals (Doloreux and Shearmur 2012). The governance of the business processes has been supported by the “know-why” knowledge component that provides situational knowledge and guidelines for making decisions faced business situations (Le Dinh et al. 2013).

Method of the BDD-KMS Architecture

A method is defined as the activities that support the process of organizational knowledge development. A BDD-KMS operates based on four main activities of the knowledge management process (Le Dinh et al. 2015): knowledge capture, organization, transfer and application. They are respectively realized by the DaaS, IaaS, KaaS and BPaaS layers as the followings:

- The DaaS layer captures batch and real-time data through the data loading and ingestion components. The batch data is loaded, processed and saved to the data repository, whereas the real-time data is ingested and sent to the real-time processing component. The data services enable maximal mashup, reuse, and sharing of the available data in the data repository.
- Based on the knowledge structure, the batch and real-time data processing components of the IaaS layer filter, extract, analyze, integrate and migrate the batch and real-time data into batch and real-time information views, respectively. The information views are unified and saved to the knowledge repository that may store its data in the data repository. Besides, the real-time component also writes the processed real-time data to the data repository. The information services make the up-to-date information available to users.
- The KaaS layer is able to deduce knowledge from the information of IaaS through the knowledge services. The collaboration service allows the creation, sharing and application of the knowledge and users with support of tools. The discovery service provides functions such as search, retrieval, mining, mapping, navigation, and presentation of the knowledge. The learning service creates more knowledge for decision-making by enabling machine learning algorithms, data scientists, predictive modelers, and other analytics professionals. The decision-making service devises an increased understanding about a business task or agreed criteria, and a guideline to make the most effective and strategic business decision. The control services ensure the high-quality performance for business processes and their corresponding services. All the knowledge generated is then saved to the knowledge repository.
- The BPaaS layer, on top of the other service layers, delivers the services to applications through combining them with knowledge-intensive business processes. The main purpose of this layer is to deal with the practices and applications to facilitate the development of intellectual capital by applying the knowledge repository to a special use or purpose. There are three primary mechanisms for knowledge application: directives, organizational routines, and self-contained task teams (Le Dinh et al. 2013).

It is obvious that the proposed architecture satisfies the two objectives of the research question. Namely, the DaaS and IaaS layers carry out the capturing and unifying knowledge discovered from big data (based on the architecture’s constructs and model). Moreover, all the layers are designed according to SOA principles to support the activities of the knowledge management process (based on the architecture’s method). As a result, the architecture can take the benefits from SOA and enhance the added value from big data in order to promote the organizational knowledge management process.

Design Evaluation

The design evaluation is based on the analytical evaluation method (Hevner et al. 2004) to evaluate the utility, quality and efficacy of the proposed architecture. For the time being, this research focuses on the utility and quality of the architecture. Therefore, this evaluation has been carried out in two steps: one for static analysis and one for architecture analysis. The first step concerns the static analysis, which
Towards an Architecture for Big Data-Driven Knowledge Management Systems

examines the artefacts for static qualities. This step has been performed by two researchers, one in information systems and one in knowledge management, which play the role of the users of the design. These users evaluate the design and then the artefacts of the design. The results of the first step have been used to design the architecture as mentioned in Figure 2 that supports all the knowledge development levels. The second step concerns the architecture analysis, which studies the suitability of the design artefacts into current technologies and tools related to big data and information system architecture. The results of the second step helped us to fulfil the architecture, which is composed of important features and viable combination of software tools to meet the requirements of the big data-driven KMS architecture.

Figure 3 presents a use case extracted from an implementation of the BDD-KMS architecture. Several big data technologies have been launched and could replace highly-customized, expensive legacy systems with a standard solution that runs on commodity hardware. Most often, the open-source technologies are prominent candidates to be used in big data projects because they can be implemented far less expensive than proprietary technologies. Besides, these technologies have a great support community as well.

Figure 3. A SO-Based Big Data-Driven KMS

Technical details for implementing the service layers of the BDD-KMSs are described as follows.

The Data Sources Layer. The batch data sources come from files, databases, or information systems such as Enterprise Content Management, Customer Relationship Management and Enterprise Resource Planning systems, as well as NoSQL-based systems. A NoSQL database provides a mechanism for storage and retrieval of data, which is different to the tabular relations used in relational databases. On the other hand, the real-time data sources are feeds, events, comments, messages, and images of social network applications, mobile devices, and sensors.

The DaaS Layer. Hadoop Distributed File System (HDFS) is selected as the data repository of the system. It is a well-known fault-tolerant distributed file system of Apache Hadoop, a dominant framework...
for big data processing with large infrastructures being deployed and used in manifold application fields (Fotaki et al. 2014). Concerning data capture tools, Apache Sqoop is widely used for efficiently exporting or importing bulk data between HDFS and structured data storages such as relational databases (Pippal et al. 2014). ETL (Extract, Transform, and Load) tools can be used to bulk load data from files or NoSQL to HDFS. On the other hand, Apache Kafka, which is a distributed, reliable, high-throughput and low latency publish-subscribe messaging system, is used for stream processing (Assunção et al. 2015). The data service is executed by WebHDFS for Hadoop that enable users to access data in HDFS using an industry-standard mechanism.

The IaaS Layer. The IaaS is implemented based on Apache Hbase and Spark. Apache HBase is a distributed column-oriented NoSQL database built on top of HDFS that supports random, real-time read/write access to HDFS, and the bulk loading feature (Assunção et al. 2015). Since HBase can provide web service gateways, this technology is used to implement the information services related to the knowledge repository. Unlike a conventional KMS architecture designed for structured and internal data, the big data-driven KMS architecture works with raw unstructured and structured data as well as internal and external data sources. Since the ability to handle batch and (near) real-time data is required, Apache Spark is selected for both the batch and real-time data processing. Spark has emerged as the next-generation big data processing engine because it works with data in memory that is faster and better to support a variety of compute-intensive tasks (Shanahan and Dai 2015). Notably, Spark’s components, including stream processing (Streaming), machine learning library (MLlib), structured data querying (SQL), and GraphX graph processing, are built on the same core architecture (Core/Engine) for distributed analytics. Spark Core works with the batch data from the data repository (HDFS) to organize content according to their semantics and to create and maintain the knowledge base (HBase).

The KaaS Layer. From a technical perspective, the system built upon SOA principles is constituted from the services as mentioned earlier. They are defined by a description language with callable interfaces to support business processes and are implemented in different programming languages. As a result, we found an ideal technology, Web services (WS) suiting to the implementation of all the knowledge services. Web services are the most popular and well-known technology to implement SOA, both in the industry and the academia (Chollet and Lalanda 2011).

The BPaaS Layer. The primary mechanisms for knowledge application can be directives and organizational routines. Directives refer to the specific set of rules developed through the conversion of tacit knowledge to explicit knowledge for efficient; meanwhile, organizational routines refer to the development of process specifications that allows individuals to apply and integrate their specialized knowledge without the need to communicate what they know to others (Le Dinh et al. 2013). It is important to model business processes and then to use the workflow application to support the automatization of business processes and the rationalization of organization routines as well as to integrate directives into processes to facilitate organizational learning and intelligent capital development.

Conclusion

In the era of big data, one of the most important characteristics of knowledge management systems (KMSs) should be big data-driven to leverage all available data as a competitive advantage. We presented a service-oriented architecture for implementing big-data driven KMSs (BDD-KMS). It is the first service-oriented architecture that concentrates on reconciliation of big data and KMSs to facilitate organizational learning based on the perspective of knowledge objects (Olivio et al. 2016).

Our approach aims at adding more business value from big data and at facilitating knowledge development and organizational learning. Big data has transformed enterprises into data-driven organizations, which require the foundation to transform data into knowledge, optimize decisions, and maximize profits. The approach helps data-driven organizations to build a new-generation of KMSs based on the service orientation that support the organizational knowledge development process and the unification of knowledge derived from diverse (big) data sources. By applying service-oriented principles, an organization can manage and govern business and digital transformation, setting them apart from their competitors. The benefits include seamless integration, cloud enabled solutions, holistic business insight and organizational agility.
In order to validate and adapt this proposal to recent IT trends, we plan to use the proposed architecture to develop and deploy the BDD-KMSs in practice. Currently, an ongoing project is being carried out to develop a prototype for customer knowledge management. This project uses the BDD-KMS approach to transform an open-source content management system into a customer KMS and then integrates this system with other information systems and services to capture and organize customer knowledge in a coherent way. This project helps us to perform other analytical design evaluation analyses such as optimization and dynamic analysis to evaluate the efficacy of the architecture (Hevner et al. 2004). In the future, the prototype can be used for other evaluation methods such as testing, experimental and observational methods.

Finally, there are two critical research directions to increase the business value of our BDD-KMS approach: i) A workflow engine for autonomous and run-time execution of business processes in the BPaaS layer, and ii) Context aware knowledge agents co-operating with this workflow engine.

REFERENCES


Towards an Architecture for Big Data-Driven Knowledge Management Systems


