Analysis of Data Warehouse Architectures: Modeling and Classification

Qishan Yang[1], Mouzhi Ge[2] and Markus Helfert[1]
[1]*Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland*
[2]*Faculty of Informatics, Masaryk University, Brno, Czech Republic*
qishan.yang@insight-centre.org, mouzhi.ge@muni.cz, markus.helfert@dcu.ie

Keywords: Data Warehouse, Architecture, Classification, Modeling, Big Data, Archimate

Abstract: With decades of development and innovation, data warehouses and their architectures have been extended to a variety of derivatives in various environments to achieve different organisations' requirements. Although there are some ad-hoc studies on data warehouse architecture (DWHA) investigations and classifications, limited research is relevant to systematically model and classify DWHAs. Especially in the big data era, data is generated explosively. More emerging architectures and technologies are leveraged to manipulate and manage big data in this domain. It is therefore valuable to revisit and investigate DWHAs with new innovations. In this paper, we collect 116 publications and model 73 disparate DWHAs using Archimate, then 9 representative DWHAs are identified and summarised into a "big picture". Furthermore, it proposes a new classification model sticking to state-of-the-art DWHAs. This model can guide researchers and practitioners to identify, analyse and compare differences and trends of DWHAs from componental and architectural perspectives.

1 INTRODUCTION

In an organisation, data usually come from heterogeneous resources that include different views and may provide inconsistent information to end users. Thus, an organisation needs a systematic channel to maintain and manipulate its data. Although operational database management systems (ODBMS) could partially solve the problem, it mainly serves for running daily business transactions, which focuses more on managing active interactions and data on time. Therefore, a Data Warehouse (DWH) is necessary to manage the business and support strategic decisions for an organisation (Ge and Helfert, 2006). (Devlin and Cote, 1996) states that a DWH is a single, complete, and consistent store of data from various sources and made available to end users in a business context. This centralised data system governs and maintains data for analyses to meet an organisation's requirements.

DWHs need to rely on architectures and other components assisting them, e.g. solving data quality issues and integrating data (Yang et al., 2017). Data warehouse architectures (DWHAs) are developed synchronously along with DWHs' progress to mainly serve for DWHs. A variety of DWHAs is proposed in order to fit different requirements and situations. To the best of knowledge, there is no classification which categories all the DWHAs. (Ariyachandra and Watson, 2008) and (Matouk and Owoc, 2012) conduct surveys on parts of DWHAs and calculate their distributions mainly in the United States and Poland respectively. They only cover five types of DWHAs. (Hub-and-Spoke, Data Mart Bus, Centralised, Independent Data Marts and Federated). The latest DWHA classification is proposed by (Blai et al., 2017), but some DWHAs, such as Virtual DWHA, Big DWHA and DWH with data lake architecture (DLA), are still excluded. These excluded DWHAs also frequently appear in literature and industry, which solve specific problems in terms of addressing big data issues (Inmon, 2016) and swift data analyses (Pasquale et al., 2008). While, the DHWAs included in their research above have few advantages to solve these problems, which would be more time-consuming and less-efficient. Hence, it is worthy to discuss and classify them together. It could benefit people having more choices to achieve their needs especially in big data and high-efficient era.

In order to systematically conduct this research, a systematical literature review is taken to collect DWHAs in literature. Due to different descriptions of DHWAs, they are generalised and modelled in order to better demonstrate and differentiate them. It further analyses the distribution of each DWHA in literature then compares the result with two previous

surveys to give insights from industrial and academic perspectives. These DWHAs identified from literature are classified for better structuring and presenting their relationships. In addition, as intrinsic features of DWHAs, some of them are analogous and difficult to distinguish. Hence, it is necessary to discuss and compare their dissimilitudes. The main contributions of this research are summarised below.

- identified a plethora of DWHAs based on a systematic literature review;
- modelled and generalised DWHAs to describe similarities and differences among them;
- conducted an analysis of DWHAs to offer an overview of their distributions;
- proposed a classification model to analyse different types of DWHAs.

The remainder of the paper is organised as follows. Section 2 reviews the related work of previous research in this domain. Section 3 presents the research methods applied in this research. Section 4 explains the results based on data from literature. Section 5 proposes a DWHA classification model and further analysis. Finally, section 6 summarises this research and outlines future work.

2 RELATED WORK

With decades of development, DWHs and their architectures are not only focusing on data analyses and decision-making, but they also follow closely behind the ever-accelerated updating of science and technology to meet new requirements. This paper mainly focuses on investigating DWHAs. As the foremost components in these architectures, DWHs also should be concerned and discussed hereby. There are different types of traditional DWHAs to address size-controlled data sets. They are extended to different types (e.g. big DWH and data lake) in terms of the big data challenge. Some relevant concepts are discussed in this section to clear the boundary of this research.

2.1 Traditional Data Warehouse

A traditional DHW herein normally refers to a DWH mainly addressing size-controlled structured data (Katal et al., 2013). Different understandings of DWHAs can be found in literature. (Inmon, 1997) defines that a DWH is a subject-oriented, integrated, non-volatile and time-variant collection of data that supports of managements decision-making capabilities. (Kimball and Ross, 2013) advocates that a DWH is the conglomerate of all DMs within the company and information is always stored in the dimensional model. Two main approaches of developing DWHs are the top-down and bottom-up (George, 2012). In the top-down approach, a centralised DWH is developed first, then building independent data marts (DMs) for individual departments. However, in the bottom-up approach, DMs are built first then the DWH is created by combining these DMs.

In addition, there is another terminology called "data warehousing", which is similar to the DWH but with a different meaning. Data warehousing more focuses on processes or mechanisms of how to organise algorithms, components and data to build data warehouses. When comparing the DWH and data warehousing, the former more focuses on storing and managing data with several characteristics mentioned above, while the data warehousing lays emphasis on the process of building a DWH project. Hence, when collecting data, this terminology is not included in this research.

2.2 Big Data Warehouse

Nowadays, DWHs are facing exploded data that is usually referred to as big data. Especially, unstructured data dominates a large portion of big data. This type of data contributes and drives the innovation which should be concerned if more data needs to be involved in the process of analysis. It constitutes 90 percent of information, which includes social text, audio, video, sensors, and click stream data, etc. While, the traditional DWH systems are not designed to scale, and handle this exponential growth of unstructured data (Pasupuleti and Purra, 2015). In order to address new unstructured big data challenges, DWHs are extended and merged with state-of-the-art technologies. The big data platforms or concepts (e.g. Hadoop and Hive) are leveraged to build DWHs with high-speed processing capacities (Thusoo et al., 2010b).

These DWHAs associated with state-of-the-art technologies are indispensable to address big data issues especially in manipulating a mass of unstructured data. For instance, (Thusoo et al., 2010b) applies Hive on top of Hadoop to develop a big DWH to store and analyse log data at Facebook. This research mainly focuses on the reconciled layer. (Yang and Helfert, 2016b) investigates and builds a data warehouse architecture in the context of big data. It is based on the Flume, Hadoop, Hive, HBase, Sqoop etc. to establish a DWHA to store and analyse website log data. In this research, the reconciled layer and derived layer are built and mixed in the Hadoop platform.

2.3 Data Lake

In addition, as a novel data repository, the Data Lake (DL) is also an option to address big data issues. According to the literature, a DL not only can be applied to manage data solely, but it can also be accompanied by a DWH. (Research, 2014) holds that the DL and the enterprise DWH provide a synergy of capabilities to deliver accelerating responses. They cooperate with each other to present their competences respectively. This combination enables people to deal with enormous data and driving business results. There are different descriptions and definitions of the DL and its architecture. James Dixon proposed that data in a DL is more natural than the data in a DM in which data is cleansed, packaged and structured for easy consumption (Dixon, 2010). A DL is a collection of various data assets which are near-exact or even exact of the source format. It serves only for the most highly skilled analysts to explore data.

A DL is thus less structured compared to a traditional and big DWHs as data preparation is eliminated. It is introduced to maintain and manipulate the explosively growing unstructured data. This novel system has abilities to capture and maintain various types of raw data with a low cost, operate transformations on the data, define the schema on the fly, perform new types of data processing, answer ad hoc analyses based on the specific use cases (Research, 2014). (Devlin, 2014) summarises that a DL is an idea that all enterprise data can and should be stored in Hadoop. It is a rip-and-replace strategy for all DWHs, DMs and eventually even operational databases because of its fullest extent.

2.4 DWHA Classifications

According to the literature, there are two main DWHA classifications which observe DWHAs from different views. The first classification includes single-layer, two-layer and three-layer architectures that are depending on the number of layers used by the architecture (Devlin and Cote, 1996; Senapati and Kumar, 2014). The second classification consists of Independent DMs architecture, Bus, Hub-and-Spoke, Centralized and Federated architectures based on components and relationships (Ariyachandra and Watson, 2008).

In terms of these three types of DWHAs, The primary mechanism of the single-layer (real-time layer) is all data sets are stockpiled once only (Inmon, 1997). An operational system and a data warehousing system share the identical data that are stored in the same place. The two-layer architecture adds a derived data layer to enable the operational and analytical requirements to be fulfilled separately. This approach addresses the single-layer architecture problems, but it also has the limitation, which has low performance in inhomogeneous or large volumes of data from different sources (Devlin and Cote, 1996). The three-layer architecture is comprised of a real-time data layer, a reconciled data layer and a derived data layer, which adds a reconciled layer as the new layer compared with the two-layer architecture. This new layer materialises the integrated and cleansed data.

When investigating DWHAs from components and their relationships, they can be mainly categorised into five predominant architectures (Ariyachandra and Watson, 2006; Ariyachandra and Watson, 2010). Although other architectures (e.g. hybrid) are mentioned in this domain, they tend to be variations on these five (Scheibe and Nilakanta, 2008). (Ariyachandra and Watson, 2008) conducts a Web-based survey Surveyed companies ranged from small (less than $10 million in revenue) to large (in excess of $10 billion). Most of them are located in the United States (60%) covering a variety of industries. (Matouk and Owoc, 2012) does a similar survey and collected data from firms listed in the Warsaw Stock Exchange. The details of these five DWHAs and the two surveys will be discussed in the following sections.

3 RESEARCH METHODOLOGY

In this paper, the quantitative research method is applied to navigate the processes and output numerical results. This method is presented as ”quantities” or numbers, which is normally generate results by doing statistical analyses (Patten and Newhart, 2018). It normally explains phenomena by analysing collected numerical data using mathematically based methods (e.g. statistics)(Aliaga and Gunderson, 2006). There are different patterns to conduct its processes. (Muijs, 2010) lists four situations of using quantitative methods. (Muijs, 2010) organises the processes of quantitative research with six steps, which is chosen and applied in this research: (1) identifying research objectives; (2) selecting research design; (3) defining what information is needed; (4) executing research instruments; (5) collecting data; (6) analysing data.

Accordingly, our research objectives are set down in the first step, which is investigating the distribution and classification of DWHAs. (Robson and McCartan, 2018) proposes three research designs: experimental, quasi-experimental and nonexperimental, which are widely accepted in different domains of research. In the second step, nonexperimental and its

sub-branch (analysis of existing data sets) are used for this research. In the third step, the information is needed in this research including publications regarding DWHAs, hence, the publications are defined with the keywords of data warehouse architecture. As for the fourth and fifth, the systematical literature review is executed as the research instrument to collect data from existing data sets. The data analyses for the last step are presented in section 4 and 5.

4 RESULTS

DWHAs are identified and generalised by analysing the contents and diagrams of these related publications collected from literature above. Different descriptions can be found to present even for the same DWHA in these publications. In order to unify them, A modelling language, Archimate, is leveraged to model and diagram their components and relationships. Archimate is the open source modelling toolkit for all levels of Enterprise Architects and Modellers (Modelling, 2018). The conformed modelled DWHAs can facilitate to observe the similarity and differentiation of DWHAs.

4.1 DWHA Overview

Archimate is applied to model 73 distinct DWHAs extracted from 116 articles. Each identified DWHA is described and explained below. At the meantime, in order to better demonstrate and distinguish the differences among them, these modelled DWHAs are generalised and merged together in figure 1.

The Hub-and-Spoke DWHA: Bill Inmon proposes the Hub-and-Spoke DWHA, because it is suitable for traditional relational database tools to fulfill the development requirements of an enterprise-wide DWH (Breslin, 2004). It is a top-down approach to develop a DWH system, which means a centralised DWH is built first, then DMs are built on this DWH.

The Data Mart Bus DWHA: Ralph Kimball (Kimball and Ross, 2013) supports the Data Mart Bus using dimensional modelling, in which a data modelling approach is unique to the DWH. The Enterprise-wide cohesion is accomplished by a data bus standard (Breslin, 2004). It is a bottom-up approach to develop a DWH. Conformed dimensional tables and DWH bus matrices play vital roles to provide consistent dimensions for implementation of DMs asynchronously.

The Centralised DHWA: The Centralised DHWA collects heterogeneous data into the

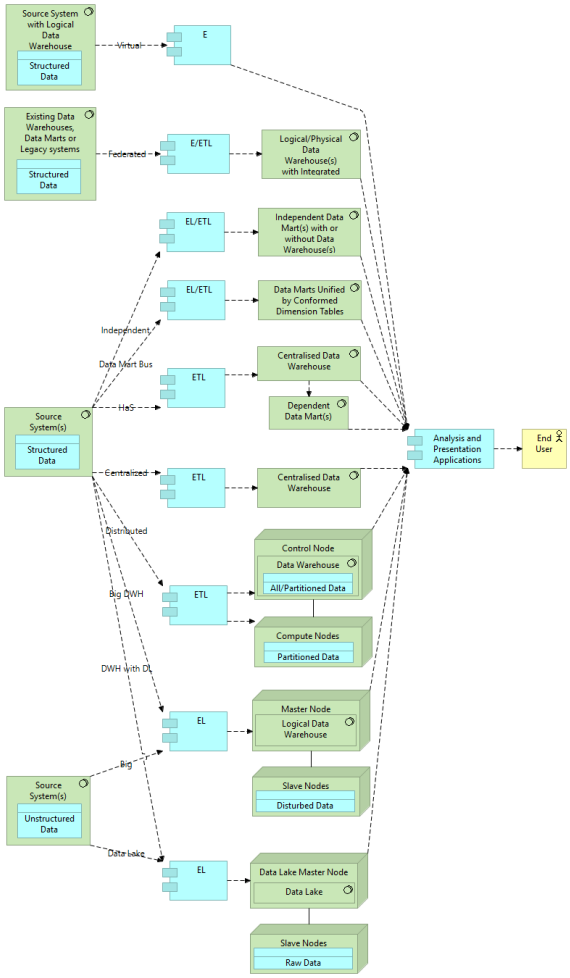

Figure 1: Data Warehouse Architecture Overview

enterprise-level cross-functional information system, which enables the separate sources to be used together and stores the presentation-ready format for requests from end users or further analysis. The topology of this DWH is not complicated to maintain warehouse data for many end-users and applications throughout the organization (Bontempo and Zagelow, 1998).

The Independent DWHA: In this research, the independent DHWA originally derives from a broader scope rather than the Independent Data Marts. The Independent Data Marts consist of physically/logically separated and irrelevant DMs. while the Independent DHWA may contain hybrid and less-coupling components including DWHs and DMs. This type of architecture may be called as the DM architecture which has the independent DM(s) or distributed DHWA that contains the mixed but independent DWH(s) and DM(s) like the architecture in (Sahama and Croll, 2007).

The Federated DWHA: The Federated DWHA is based on the bottom-up implementation, in which enterprise-level dependent DMs can provide the enterprise-wide analytical solutions. The federated DWH model can be acted as the added data staging layer that enables easier implementation of analytical solutions to hide the complexities of operational data sources (Alsqour et al., 2012).

The Virtual DWHA: The virtual DHW manipulates and analyses operational data sources with limited data integration functions and also summary views of the Data Warehouse depending on the complexity of requirements (Chaki and Sarkar, 2010). Middleware may be leveraged as an interface between front-end tools and the operational database (Ayman et al., 2001). (Devlin and Cote, 1996) describes that a virtual data warehouse is a logical warehouse allowing users to directly access multiple operational data through middleware tools. It does not necessarily store copies of the operational data with different formats compared with other DWHAs.

The Distributed DWHA: The distributed DWHA herein mainly includes DWHs working in parallel or built on multi-node cloud computing platforms, in which data are pre-processed and physically distributed in ODBMS (e.g Oracle, DB2 etc.) with a predefined schema. This architecture fragments the warehouse schema using partitioning algorithms (Pavlo et al., 2012) and allocates the generated fragments over the compute nodes using particular algorithms(Menon, 2005). In order to gain availability of data and the high performance of the system, the generated fragments may be duplicated and saved in different compute nodes.

The Big DWHA: This DWH is built on a big data platform (e.g. Hadoop), in which the distributed file system and other mechanisms (e.g. MapReduce) are applied to store and deal with data respectively (Yang and Helfert, 2016b). This DWH can address big data issues (e.g. petabyte level) for reporting and ad-hoc analyses by using SQL-like language (e.g Hive) (Thusoo et al., 2010a), which is easily executed by people with SQL experience. It can manage unstructured data provided by a logical DWH (Yang and Helfert, 2016a). This logical DWH does not need to strictly pre-process data with a predefined schema.

The Data Lake Architecture: The DL can be leveraged directly as a centralised raw data repository providing information for further analyses (Rangarajan et al., 2015). The DL and DWH can also cooperate together to address the big data issues and achieve data analysis requirements, which are generalised and modelled in figure 1. (Inmon, 2016) presents an architecture with a DWH and DL. In their architecture, the DWH populates the DL, which means the DWH is built first or a company already has a legacy DWH then build a DL. Another situation is the DWH fed by the DL (Pasupuleti and Purra, 2015). The DWH obtain data from the DL, which is suitable for building the DL then the DWH, optimising the legacy DWH data flows or extending the DWHA to meeting intangible requirements.

4.2 DWHA Distribution Analysis

A statistics is made to analyse the distribution of these DWHAs in literature and compare with the two previous research (Ariyachandra and Watson, 2008) and (Matouk and Owoc, 2012) presented in table 1. There is no big conflict among these three results. The first three DWHAs (Hub-and-Spoke, Data Mart Bus and Centralised) are similar. In the first two research, the ratios of Data Mart Bus are greater than the Centralised DHWA's, while in this paper, the result is the opposite. The reason might be observing the domain from different angles (e.g. industrial or academic). Therefore, the three high-weighed DWHAs are widely accepted in practice and academia. When choosing or building DWHAs, these three architectures should be concerned. In addition, most of the papers about DLAs have been published since 2015, so it shares less proportion. If putting the distributed DWHA, Big DWHA and DLA together, these three architectures' occupation cannot be neglected, which may imply that big data issues are increasingly inescapable in the DWHA context. Hence, when designing and developing DWHAs, it is necessary to consider these three big-data-oriented architectures.

5 DWHA CLASSIFICATION MODEL

As discussed before, based on literature, there are two main DWHA classifications. The layer-based classification is general and abstract, which gives an outline of architectures. The component-based classification is derived from use cases in practice. However, they are some inter-relations between them. Some DWHAs (e.g. Hub-and-Spoke, Data Mart Bus and Centralised) may need an ETL tool to extract, transform and load data. Some DHWAs (e.g. Federated data warehouse architecture) may do not need ETL tools/staging areas, while, in order to build the logical/physical Federated DHW, the data in its underlying DWHA or DM may also need to be reconciled to provide cleansed, integrated information. Therefore, the four architectures (the Hub-and-Spoke, Data

Table 1: The data warehouse architecture distribution

	DWH Architecture Distribution								
	Hub and Spoke	Data Mart Bus	Centralised	Independent	Federated	Virtual	Distributed	Big DWHA	Data Lake
(Ariyachandra and Watson, 2008)	39%	26%	17%	12%	4%				
(Matouk and Owoc, 2012)	35%	23%	20%	15%	7%				
This paper	29%	9.3%	27%	8.7%	6.4%	2.9%	8.7%	3.5%	5.2%

Mart Bus, Centralised and Federated) have a similar structure as the three-layered DWHA has. The independent DWHA extended from the concept of independent DMs has different situations: If it consists of independent DMs mainly for individual department applications, then it may belong to traditional two-layered DWHA; If it is mixed with independent DWHs and DMs, it could belong to the three-layered DWHA as it may also need to reconcile data from different sources. Therefore, the relationship between these two main classifications (Layer Based Classification and Component Based Classification) can be illustrated: Two-Layer $\supset$ {Independent [DM]}; Three-Layer $\supset$ {Hub-and-Spoke, Data Mart Bus, Centralised, Independent [DWH with DM] and Federated}. By analysing the two main DWHA classifications and collected DHWAs from literature, we propose a new classification model of DWHAs based on their characteristics, which are structured in the top of table 2 with bold font.

5.1 Discussions and Implications

Even DWHAs are classified in this model, there are still some indistinct characteristics or similar components which could confuse their differentiation. In order to easily identify and explain the differences of these DWHAs, three features and fifteen sub-features are chosen and discussed, which mainly derive from generalised and modelled DWHAs in figure 1. Some typical DHWAs collected from literature are selected and discussed here as cases. The table 2 analyses and summarises differences of these cases. The tick in this table means that a certain DWHA has this sub-feature or can address this sub-feature. For example, the Virtual DWHA manipulates structured data and then a tick is placed in the related cell. As can be seen in this table, differences could be analysed from the feature perspective. For example, most of DWHAs cannot directly process unstructured data except Big DWH and DL. Hence, if a company needs to analyse a large amount of unstructured data directly, these two architectures may be more suitable than others. In order to distinguish the differences and better understand these DWHAs, some of them are analysed together, as they have similar components or in the same branches.

Virtual, Federated and Big Data: From figure 1, the Virtual, Federated and Big DWHAs could contain a logical DWH, but from table 2, the virtual DWHA normally manages data with a single data schema in a source system, while, the Federated and Big DWHAs could extract data from multiple data sources. In addition, the Big DWHA store data using the distributed file system (DFS); the Federated DWHA may extract data from different sources and save data in traditional database management systems.

Independent [DWH with DM], Centralised, Data Mart Bus, Hub-and-Spoke and Federated: In the Classification Model, the Centralised, Data Mart Bus, Hub and Spoke, and Federated DWHAs are categorised in the three-layer branch. The differences mainly can be discerned in the ETL and DM features. For example, the Data Mart Bus has separated DMs but connected with conformed dimension tables; the Hub-and-Spoke has DM(s) getting data from its depended DWH, but they do not need to be strictly linked by conformed dimension tables. The Independent [DWH with DM] has DWH(s) and DM(s), but they are independent.

Distributed, Big DWHA and Data Lake: The Distributed, Big DWHA and DL belong to the same main branch. The data could be extracted, loaded then transformed (ELT) in the Big DWHA , as the Big DWHA has powerful computing platforms (e.g. Hadoop) to manipulate data. In addition, its big DWH (e.g. Hive) does not necessarily pre-load data, which only needs to be informed where the data is stored in DFS. However, data in the Distributed DHW should be extracted, transformed then loaded into a DHW with a predefined schema (ETL). Besides, the DL stores row data with detailed granularity, which extracts and loads (EL) row data from sources with less pre-processing.

Table 2: DWHA Classification Model

DWHA Classification Model		Data Warehouse Architecture									
		Traditional Data Oriented Architecture							Big Data Oriented Architecture		
		One Layer	Two Layer	Three Layer					Big		Synergetic
		Virtual	Independent [DM]	Independent [DWH with DM]	Centralised	Data Mart Bus	Hub-and-Spoke	Federated	Distributed	Big DWHA	Data Lake
Data Type	Structured Data	✓	✓	✓	✓	✓	✓	✓	✓		
	Structured & Unstructured Data									✓	✓
Schema	Single	✓	✓								
	Multiple			✓	✓	✓	✓	✓	✓	✓	✓
ETL	Extract	✓						✓			
	Extract & Load		✓			✓					✓
	Extract & Transform										
	Extract & Load & Transform									✓	
	Extract & Transform & Load			✓	✓		✓		✓		
Data Mart	Conformed Data Mart					✓					
	Independent Data Mart		✓	✓							
	Dependent Data Mart						✓				
	Unknown or Null	✓			✓			✓	✓	✓	✓
Storage	Distributed								✓	✓	✓
	Aggregated	✓	✓	✓	✓	✓	✓	✓			

6 CONCLUSION

This paper conducts a quantitative study associated with a systematical literature review on DWHAs to investigate them from the traditional data oriented and big data oriented perspectives. 116 publications are collected from literature and 73 disparate DHWAs are modelled using Archimate. Based on these modelled DWHAs and their characteristics, 9 representative DWHAs are identified and summarised into a "big picture" which directly demonstrates their similarities and differences. The number of these architectures are counted and their distribution is generated by the statistic technique. Furthermore, a new DWHA classification model is proposed to better identify, analyse and compare DWHAs from featured and structural perspectives. For future work, the classification model would be refined with more features to better measure the performance or difference of these DWHAs. Also, this model would be extended to accommodate new technologies and DWHAs in further research.

ACKNOWLEDGEMENTS

This publication is supported by the Science Foundation Ireland grant SFI/12/RC/2289 to Insight Centre for Data Analytics. INSIGHT is funded under the SFI Research Centres Programme and is co-funded under the European Regional Development Fund.

REFERENCES

Aliaga, M. and Gunderson, B. (2006). *Interactive Statistics, 3rd Edition*. Pearson.

Alsqour, M., K, M., and Owoc, M. L. (2012). A survey of data warehouse architectures: preliminary results. In *InFedCSIS (pp. 1121-1126)*.

Ariyachandra, T. and Watson, H. (2010). Key organizational factors in data warehouse architecture selection. In *Decision support systems. 2010 31;49(2):200-12*.

Ariyachandra, T. and Watson, H. J. (2006). Which data warehouse architecture is most successful? In *Business intelligence journal, 11(1), p.4*.

Ariyachandra, T. and Watson, H. J. (2008). Which data warehouse architecture is the best? In *Communications of the ACM, 51(10), p.146*. ACM.

Ayman, A., Zaiane Osmar, R., and Ji, Y. (2001). Towards framework for the virtual data warehouse. In *British National conference on databases (pp. 202-218)*.

Blai, G., Poi, P., and Jaki, D. (2017). Data warehouse architecture classification. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on (pp. 1491-1495)*. IEEE.

Bontempo, C. and Zagelow, G. (1998). The ibm data warehouse architecture. In *Communications of the ACM, 1;41(9):38-48*. ACM.

Breslin, M. (2004). Data warehousing battle of the giants. In *Business Intelligence Journal*.

Chaki, N. and Sarkar, B. (2010). Virtual data warehouse modeling using petri nets for distributed decision making. In *Journal of Convergence Information Technology, 5(5), pp.8-21*.

Devlin, B. (2014). Data lake muddies the waters on big data management. In *Available at: https://searchbusinessanalytics.techtarget.com/feature/Data-lake-muddies-the-waters-on-big-dat-management [Accessed 15 Novmber 2018]*.

Devlin, B. and Cote, L. (1996). *Data warehouse: from architecture to implementation*. Addison-Wesley Longman Publishing Co., Inc.,.

Dixon, J. (2010). Pentaho, hadoop and data lakes. In *Available at: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ [Accessed 15 Novmber 2018]*.

Ge, M. and Helfert, M. (2006). A framework to assess decision quality using information quality dimensions. In *11th International Conference on Information Quality*, pages 455–466.

George, S. (2012). Inmon vs. kimball: Which approach is suitable for your data warehouse. data warehous. In *pp.1-12*.

Inmon, B. (2016). *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics.

Inmon, W. (1997). What is a data warehouse? In *Prism Tech. Topic 1(1)*.

Katal, A., Wazid, M., and Goudar, R. (2013). Big data: issues, challenges, tools and good practices. In *Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409)*. IEEE.

Kimball, R. and Ross, M. (2013). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Matouk, K. and Owoc, M. L. (2012). A survey of data warehouse architecturespreliminary results. In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on (pp. 1121- 1126)*. IEEE.

Menon, S. (2005). Allocating fragments in distributed databases. In *IEEE transactions on parallel and distributed systems, 16(7), pp.577-585*.

Modelling, A. (2018). Archi open source archimate modelling. In *Available at: https://www.archimatetool.com/ [Accessed 15 Novmber 2018]*.

Muijs, D. (2010). *Doing quantitative research in education with SPSS*. SAGE Publications Ltd.

Pasquale, C., Ionescu, B., Ionescu, D., and Gurerero, A. (2008). Virtual data warehouse architecture for real-time webgis. In *Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. IEEE Conference on*. IEEE.

Pasupuleti, P. and Purra, B. (2015). *Data Lake Development with Big Data*. Packt Publishing Ltd.

Patten, M. L. and Newhart, M. (2018). *Understanding Research Methods: An Overview of the Essentials*. Taylor & Francis, New York and London.

Pavlo, A., Curino, C., and Zdonik, S. (2012). Skew-aware automatic database partitioning in shared-nothing, parallel oltp systems. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 61-72)*. ACM.

Rangarajan, S., Liu, H., Wang, H., and Wang, C. (2015). Scalable architecture for personalized healthcare service recommendation using big data lake. In *Service Research and Innovation (pp. 65-79)*. Springer, Cham.

Research, C. (2014). Putting the data lake to work a guide to best practices. In *Teradata*.

Robson, C. and McCartan, K. (2018). *Real world research*. John Wiley & Sons.

Sahama, T. and Croll, P. (2007). A data warehouse architecture for clinical data warehousing. In *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*. Australian Computer Society, Inc.

Scheibe, K. and Nilakanta, S. (2008). Dimensional issues in agricultural data warehouse designs. In *pp.263-278. electronics in agriculture, 60(2)*.

Senapati, R. and Kumar, D. A. (2014). A survey on data warehouse architecture. In *International Journal of Innovative Research in Computer and Communication Engineering.*

Thusoo, A., Sarma, J., Jain, N., Shao, Z., Chakka, P., Zhang, N., Antony, S., Liu, H., and Murthy, R. (2010a). Hive-a petabyte scale data warehouse using hadoop. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on (pp. 996-1005)*. IEEE.

Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., Sarma, J. S., Murthy, R., and Liu, H. (2010b). Data warehousing and analytics infrastructure at facebook. In *Proceedings of the 2010 ACM International Conference on Management of data (pp.1013-1020)*.

Yang, Q., Ge, M., and Helfert, M. (2017). Data quality problems in tpc-di based data integration processes. In *International Conference on Enterprise Information Systems (pp. 57-73)*. Springer, Cham.

Yang, Q. and Helfert, M. (2016a). Data quality for web log data using a hadoop environment. In *ICIQ 2016: 199-208*. MIT Information Quality (MITIQ) Program.

Yang, Q. and Helfert, M. (2016b). Revisiting arguments for a three layered data warehousing architecture in the context of the hadoop platform. In *Proceedings of the 6th International Conference on Cloud Computing and Services Science, Volume 2*. SCITEPRESS.