Semantic Web 0 (0) 1
IOS Press

Semantic Modeling for Engineering Data Analytics Solutions

Madhushi Bandara [a,*], Fethi A. Rabhi [a]
[a] *Department of Computer Science and Engineering, University of New South Wales, NSW, Australia*
E-mail: k.bandara@unsw.edu.au

Abstract. Data analytics solution engineering often involves multiple tasks from data exploration to result presentation which are applied in various contexts and on different datasets. Semantic modeling based on the open world assumption supports flexible modeling of linked knowledge. The objective of this paper is to review existing techniques that leverage semantic web technologies to tackle challenges such as heterogeneity and changing requirements in data analytics solution engineering. We explore the application scope of those techniques, the different types of semantic concepts they use and the role these concepts play during the analytics solution development process. To gather evidence for the study we performed a systematic mapping study by identifying and reviewing 82 papers that incorporate semantic models in engineering data analytics solutions. One of the paper's findings is that existing models can be classified within four types of knowledge spheres: domain knowledge, analytics knowledge, services and user intentions. Another finding is to show how this knowledge is used in literature to enhance different tasks within the analytics process. We conclude our study by discussing limitations of the existing body of research, showcasing the potential of semantic modeling to enhance data analytics solutions and discussing the possibility of leveraging ontologies for effective end-to-end data analytics solution engineering.

Keywords: Semantic Modeling, Data Analytics, Software Development, Ontology

1. Introduction

The business intelligence and analytics fields are rapidly expanding across all industry sectors and many organizations are trying to make analytics an integral part of everyday decision making [1, 2]. These fields include the techniques, technologies, systems, practices, methodologies and applications that are concerned with analyzing critical business data to help an enterprise better understand its business and market and make timely business decisions [2, 3].

There is no universally accepted definition for data analytics process. CRISP-DM [11] and KD process proposed by Fayyad et. al [12] are two examples of well developed and popular definitions. In the context of this paper, we identify a "data analytics process" (also called an "analytics pipeline") as an end-to-end Data Analytics Solution (DAS) that capture tasks related to data mining, knowledge discovery or business intelligence. A Data Analytics Solution can be itself decomposed into multiple tasks such as identifying suitable datasets, developing analytics models and validation and interpretation of final results. The process that represents data analytics solution is related to the discipline of data science. The software engineering aspect of DAS, which we identify as "DAS Engineering", involves designing and developing data analytics solutions from requirement engineering to the deployment of the final solution including tasks such as requirement elicitation, data integration and process composition [13].

With the increasing popularity of big data as a research area, focus of the most research efforts have been on developing specific analysis techniques (e.g. machine learning algorithm design) but not on supporting the overall DAS engineering. Within many organizations, analysts with limited programming experience are often required to manually establish relationships between software components of the analytics solution like software services used for computation, and data

*Corresponding author. E-mail: k.bandara@unsw.edu.au.

elements or data mining algorithms [4, 5]. According to No-Free-Lunch theorem [6], the DAS engineering becomes further challenging as there is no one model that works best for every problem and depending on the application context and input data, analysts have to try different techniques before getting optimal results. Most organizations are looking for flexible solutions that align with their specific objectives and IT infrastructures [7], usually resulting in the use of a mix of data sources and software frameworks. Understanding, and managing these heterogeneous technologies needs to be supported by good knowledge management infrastructure. In addition to incurring high software development costs, maintaining and evolving heterogeneous software infrastructures in the face of constant changes in both business requirements and technical specifications is very expensive [10].

An ontology is the formal foundation for semantic modeling. The main role of an ontology is to capture domain knowledge, to evaluate constraints over domain data and to guide domain model engineering [16]. It is a powerful tool for modeling and reasoning [7]. As ontologies provide a sound representation of concepts and the relationship between concepts, they represent malleable models that are suitable for tracking various kinds of software development artifacts ranging from requirements to implementation code [17]. Such properties can provide multiple benefits to the organization such as reducing the cost of data analytics solution engineering by supporting the management of heterogeneous services, datasets, analytics models, domain knowledge, and continuously changing requirements. There has been many recent efforts in applying semantic modeling for DAS engineering, but overall picture of their capabilities is far from clear.

Hence, this study systematically explores the different research studies that are focusing on designing and developing applications that support DASs with the aid of semantic technology. Further we identify unresolved challenges and potential research directions in the analytics solution development space. We follow the systematic mapping study process proposed by Petersen et. al. [18], collect evidence from the publications in five prominent databases and extend the evidence further by snowballing [19] relevant references of identified studies. We conduct our study around the main research question of identifying the existing techniques that use semantic models in DAS engineering and two sub-questions related to that. We evaluate how different knowledge areas related to analytics solutions such as mental models of the end-user, domain knowledge, semantics of data, applicability of analytics algorithms and tools for a particular task, compatibility between data and tools etc. are represented by semantic models and leveraged for conducting tasks related to DAS engineering.

The rest of the paper is structured as follows. Section 2 describes some background related to this paper. Section 3 presents the review method that we followed. The results derived from the 82 identified studies are included in section 4 followed by a discussion in sections 5. The limitations of the conducted study are discussed in section 6 and the paper concludes in section 7.

2. Background

The literature emphasizes the significance of knowledge management in different fields such as enterprise data analytics [21] and scientific workflow [22] and there has been many attempts at identifying knowledge specific to DAS. For example, the ADAGE framework [20] proposes an approach that leverages the capabilities of service-oriented architectures and scientific workflow management systems. The main idea is that the models used by analysts (i.e. workflow, service, and data models) contain concise information and instructions that can be viewed as an accurate record of the analytics process. These models can become a useful artifact for provenance tracking and ensure reproducibility of such analytics processes. However, designing a models that accurately represent the complex business contexts and expertise associated with an analytics solution still remains a challenge. Development methods such as CRISP-DM for enterprise-level data mining [11] and Domain-oriented data mining [23] are advocating the necessity of using knowledge management techniques for capturing the business domain and understanding of data in order to build better analytics solutions.

There have been multiple known knowledge representation approaches related to different aspects of data analytics such as UML diagrams [4, 24, 25], petrinets [26] and decision modeler [21] but the focus of this paper is on semantic models that originate from the Semantic Web concept [27], where ontologies expressed via RDF[1], RDFS[2] and OWL[3] are the founda-

[1] https://www.w3.org/RDF/
[2] https://www.w3.org/TR/rdf-schema/
[3] https://www.w3.org/OWL/

tion of modeling. Although semantic technology has been part of the research landscape for a while, the industry is only just beginning to discover the power of linked open data, ontologies and semantic applications in assisting enhancements to the data analytics process. Whilst there are significant examples of leading internet companies (e.g. Google, Amazon, and Facebook) beginning to exploit the power of semantic search and domain ontologies (e.g. Schema.org, DBpedia) [28], many organizations are still largely unaware of the value that these approaches represent [29–31].

To our knowledge, there is no formal study conducted on how semantic modeling has contributed to DAS engineering except the surveys conducted by Abello et. al [7] and Ristoski and Paulheim [28]. Abello et. al [7] study is specially about using semantic web technologies for Exploratory OLAP, considering the data extraction and integration aspects. Ristoski and Paulheim [28] conducted a survey in 2016 about the different stages of the knowledge discovery process that use semantic web data. In comparison, our work is unique as it is looking at the applications of semantic models in the DAS engineering from a data analytics as well as a software engineering perspective.

3. Research Method

3.1. Introduction

As our objective was to provide an overview of how semantic technology is used in DAS engineering we conducted a systematic mapping study (SMS) process. This provides an overview showing the type of research and results that have been published by categorizing them with the goal of answering a specific research question [18, 32]. We followed the process proposed by Petersen et. al [18] to ensure the accuracy and the quality of the outcome. We conducted initial evidence search on five databases, collection of two conference proceedings and one journal related to semantic web technologies. Findings were extended through snowballing approach proposed by Wohlin [19]. The goal of our study is to present a holistic view on the use of semantic modeling in the data analytics landscape.

3.2. Research Questions

The primary focus of our SMS is to identify and understand how semantic modeling approaches are used to represent and communicate knowledge of a data analyst as well as how existing DAS engineering techniques are leveraging this knowledge. The review was conducted on a primary research question and two sub-questions which are stated as follows:

Primary Question: What are the existing techniques that use semantic modeling for data analytics solution engineering?

Sub-questions:

1. What type of concepts are modeled/used by these techniques?
2. What tasks related to DAS engineering are enabled using the identified concepts?

3.3. Search of Relevant Literature

We adapted the work used in [18, 32–34] and identified the following strategy to construct the search strings:

- Derive major terms used in the review questions
- Search for synonyms and alternative words.
- Use the Boolean OR to incorporate alternative spellings and synonyms
- Use the Boolean AND to link the major terms

To obtain a balance between sensitivity and specificity as highlighted by Petticrew and Robert [35], we selected a search string that contains three major terms related to the concepts: semantic technology, data analytics, and software engineering, connected by a Boolean AND operation. Each term contains a set of keywords related to the respective concept, connected by a Boolean OR operation.

The complete search string initially used for the searching of the literature was as follows:

(("knowledge management" OR semantic OR "linked Data" OR ontology OR "conceptual modeling") AND ("big data analytics" OR "business analytics" OR "data analytics" OR "scientific workflow" OR "data mining") AND (requirement OR "development process" OR "code generation"))

The primary search process involved the use of 5 online databases: Web of Science, Scopus, ACM Digital Library, IEEE Xplore, and ProQuest. The selection of databases was based on our knowledge about those that index major publications related to computer science, engineering and semantic technology. Based on the recommendations of domain experts, we expanded the search space to the collection of proceedings of In-

ternational Semantic Web Conference and European Semantic Web Conference, with their associated workshops, and the publications by the Semantic Web Journal accessible via DBLP's search API.

Upon completion of the primary search phase, the identification of relevant literature continued through snowballing - all the references in the papers identified from the primary sources were reviewed for relevancy. If a paper satisfied the selection criteria, it was added to the list of studies qualified for the synthesis.

3.4. Selection of Studies

Below are our exclusion criteria which were adapted from [36]:

1. Books and News Articles
2. Papers where semantic modeling was not applied directly to DAS engineering
3. Vision papers
4. Papers not written in English.
5. Application specific research that does not generalize (such as text extraction and web search applications)
6. Infrastructure related performance-oriented applications supporting distributed storage etc.
7. Full text that was not available for public access and not licensed by the University of New South Wales digital library

We did not restrict the search to a span of time. This search included all research available in the selected databases up to 27/06/2018.

3.5. Study Quality Assessment

We designed a quality checklist to measure the quality of the primary studies by reusing some of the questions proposed in the literature [35, 37]. Our quality checklist comprised 4 general questions stated below:

1. Was the study related to DAS engineering?
2. Do the studies leverage semantic models for information modeling?
3. Do they provide sound evaluation?
4. Were the findings credible?

Initially, one author went through the title, abstract and keywords of search results and divided papers into 3 categories by relevancy: "Yes", "No" and "Maybe". Then the second author went through the full text of the papers under the "Maybe" category to identify whether they were compliant or not with our quality checklist.

Through the initial database search, we identified 1414 empirical studies as candidates. Among those results, 63 (4.46% of 1414 studies) were identified as relevant studies, based on the study quality assessment and exclusion criteria. The same steps were applied to the literature identified through snowballing at the second stage as well. We iterated through the references of 63 papers selected during the initial search and identified 19 additional relevant papers for our study.

To avoid the inclusion of duplicate studies which would inevitably bias the result of the synthesis [38], we thoroughly checked if very similar studies were published in more than one paper. In total, 82 studies were included in the synthesis of evidence.

3.6. Constructing Classification Schemas

Our study requires two classification schemas to answer sub research question 1 and 2- what are the different types of semantic concepts used by identified studies and what tasks related to DAS engineering were conducted using the identified concepts.

To construct each classification schema for our mapping study we adapted the systematic process proposed in [18]. We created the classification schema in top-down fashion, incorporating different classifications proposed and used in literature to guide the classification schema construction. We used the abstract, introduction and conclusion of the selected 82 studies and aligned the studies with categories identified in literature. When necessary, the classification schema was extended with keywords and categories defined in the identified literature to provide clarity and granularity to the finalized the schema.

To answer sub-question 1, we distinguished four broad classes of concepts represented through ontologies in identified studies, referred to as domain, analytic, service and intent. This classification was guided by the proposal of Nigro [39] to use three ontology types in data mining. The first two: "Domain Ontologies" and "Ontologies for Data Mining Process" were included in our schema as the domain and analytics concepts respectively . The third one - "Metadata Ontologies" defines how variables are constructed. Because this definition is very high-level and vague we introduced two new concept types: Intent concepts and Service concepts to capture knowledge that supports requirements management and implementation management within DAS engineering. Using the evidence

of identified literature we extended this classification to a more granular level with different subtypes. The details of this classification are discussed in section 4.2.

When we look at the problem of classifying different DAS related tasks for sub-question 2, there was no unique definition in the literature regarding what constitutes a task in relation to DAS engineering. Fayyad et. al. [12] propose a five-step process model for knowledge discovery - selection, preprocessing, transformation, data mining, evaluation and interpretation. CRISP-DM proposed by Chapman et. al. [11] is more enterprise oriented and breaks down the life-cycle into five steps: domain understanding, data understanding, data preparation, modeling, evaluation, and deployment. All identified studies do not follow any specific model, some of them focus more on high level tasks such as domain understanding and process composition while others support more granular tasks such as model selection.

Hence, we combined key categories extracted from identified 82 studies with the tasks proposed in different literature. We initially identified five tasks aligning CRISP-DM process model with the tasks proposed in 82 studies - Business Understanding, Data Extraction and Transformation, Model Selection, Model Building, Result Presentation and Interpretation. We further identified four more task categories from the studies that were rooted in the software engineering literature: Data Integration, Service Composition, Analytics Solution Validation and Code Generation. These are important because they contribute to enhancing the overall quality of the DAS implementation. The identified task categories are illustrated in Fig. 1. Each DAS engineering project does not need to complete all these tasks. The selection of tasks depends on the nature of the analysis being performed (e.g. whether we need to choose amongst multiple competing data mining algorithms or use a specific algorithm) and the context in which it is being developed (e.g. whether we need to support automatic code generation to save software development cost or not).

3.7. Data Extraction and Mapping of Studies

During the data extraction phase, we read, sorted papers in accordance with the classification schemas and then reviewed them in detail. One author read and classified the 82 papers according to the two schemas, noting down the rationale of why each paper belongs to the selected category. The second author reviewed the table, discussed and resolved disagreements and compiled the final mapping. The classification schemas developed initially evolved through this phase which resulted in adding new subcategories and splitting in certain scenarios.

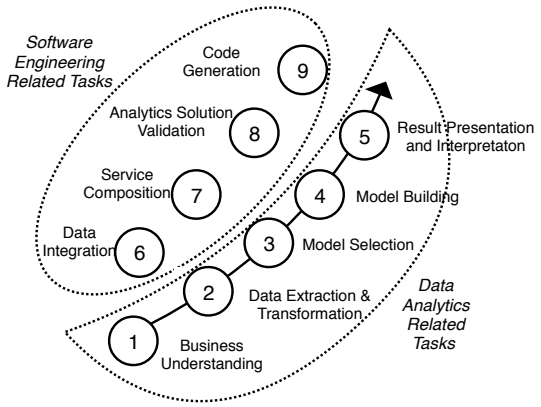

Fig. 1. Identified Tasks from Literature

The finalized mappings and the associated details are discussed in the next section.

4. Results

4.1. Primary Question: What are the existing techniques that use semantic modeling for data analytics solution engineering?

The complete list of identified studies is included in section 8. RDF, RDFS and OWL are the core building blocks of an ontology. All the identified studies, except S35 are relying on this notation for their semantic model representation (occasionally combined with other notations). S35 deviates from that common practice and only uses the Predictive Model Markup Language (PMML) [40] and Background Knowledge Exchange Format (BKEF) [41] to represent knowledge associated with a data analytics solution.

By assessing the identified studies, we observed that these efforts vary in the application context they are addressing and in the way the analytics knowledge is modeled. Through the sub-questions in section 4.2 and 4.3, we explore the different semantic concepts and tasks used by these 82 identified studies according to the classification schemas mentioned in section 3.6. Section 4.2 classifies different ontological concepts into four types and describes the characteristics

of the knowledge they capture. In section 4.3, we relate identified semantic concepts to their role in realizing and facilitating various tasks in DAS engineering.

4.2. Sub-question 1. What type of concepts are modeled/used by these techniques?

The mapping results according to the classification schema described in section 3.6 are illustrated in Table 1. We identified that out of 82 studies, the majority (54 studies) model domain concepts related to various application domains and 11 of them reuse standard domain ontologies. There are 30 studies that capture and use analytics concepts and 30 studies that capture service concepts. Smallest representation was in the intent category with 14 studies. A detailed analysis is presented in the following subsections.

4.2.1. Domain Concepts

These are context specific and high-level concepts which represent domain specific knowledge and objects. We identified 54 studies that rely on different subtypes of domain concepts.

First subtype, *application specific concepts* represent objects and relationships in a variety of niche areas such as gene and protein analysis, health-care, transport as shown in Table 1. The solutions they provide are highly coupled with a single application context and provide less opportunity for adoption by other applications. The majority of studies in this subtype (32 studies) do not propose any specific domain concepts but provide users with the ability to customize concepts in any application context.

The second subtype represents concepts that are associated with different *standard domain ontologies*. We identified five standard domain specific ontologies used for designing analytics solutions- SSN Ontology [42], GeoVocab [46], Gene Ontology [43], GALEN ontology [44] and TOVE ontology [45].

4.2.2. Analytics Concepts

Analytics concepts are closely aligned with the knowledge that reflects different algorithms, computational models and the data-flow nature of the analytics process in terms of inputs, outputs, and their compatibility. Analytic concepts provide a vocabulary for defining and communicating analytics operations and related attributes. Further, they can help in describing dependency relationships between variables. These concepts are not coupled to a specific application domain, context or implementation.

There were 30 studies that leverage analytics concepts. We classified them into three subtypes according to their role in representing different analytics tasks or methods: concepts related to data preprocessing and integration activities, concepts that capture data analytics and mining techniques and concepts that represent the control and data flow nature of an analytics process.

Under the first subtype, *data preprocessing and integration*, S14 proposes an analytics ontology for linking different temporal and geographical datasets using concepts from different predefined ontologies. S20 proposes a Rules ontology to store concepts related to rules that transform one data schema to another. S27, S37, S41 and S54 propose concepts to model different data preprocessing tasks such as null value removal, data format conversion, sampling and feature selection. S80 proposes the Analytics-Aware Ontology to represent data aggregation and comparison functions.

There are 25 studies under the subtype *data analytics and mining*. The majority of studies [S27, S36, S37, S38, S41, S53, S54, S60, S62, S66, S75, S76, S78, S79] model analytics tasks and algorithms such as classification, clustering and regression. Analytic Ontology [S15] is dedicated to statistical and machine learning models. The Actor Ontology [S24] is limited to image processing algorithms such as image classification and feature extraction. The Data Mining Ontology in S17, the Simple Data Mining Ontology in S22 and Expose ontology in S56 try to capture non-functional attributes, performance assessments such as sensitivity, specificity, accuracy and user satisfaction related to data mining algorithms. S63 and S74 propose Feature Ontologies to enhance feature set descriptions in data analytics. S67 captures event-condition-action rules for event processing vial Rule Management. S69 proposes an extended event ontology to capture event processing queries. In addition to the ontologies, three of the identified studies [S22,S27,S35,S66] capture details of the analytics models using Predictive Model Markup Language [40].

Six studies focus on *control and data flow* related concepts. The Analytic Ontology [S1], DM Workflow Ontology [S75], SemNExt Ontology [S76] and DM OPtimization Ontology [S78] include concepts that capture the data flow nature of any analytics operation with respect to an array of characteristics such as input requirements and preferences, input and output data types and accuracy. The Task-Method Ontology [S3] represents concepts that are useful in controlling and

Table 1
Classification of Semantic Concepts Modeled and Used in Identified Literature.

Classification Criteria				Study
Domain Concepts	Application Specific	Gene and Protein Analysis		S11, S27, S29, S76
		Health Care		S2, S50
		Sensor & Event Information		S30, S31, S33, S55, S69, S70, S74
		Spatiotemporal Information		S40
		Traffic Information		S18, S31, S52
		Enterprise Quality Management		S7
		Agriculture		S32
		Hyper-spectral Image Data		S24
		Power Grid		S58,S64
		Custom Built		S5, S8, S10, S12, S13, S19, S21, S23, S34, S35, S39, S42, S43, S44, S45, S47, S48, S49, S51, S57, S59, S61, S62, S63, S65, S67, S68, S72, S73, S77, S80, S81, S82
	Standard Ontology	SSN Ontology [42]		S30, S31, S33, S40, S69, S74
		Gene Ontology [43]		S11, S44, S76
		GALEN Ontology [44]		S44
		TOVE Ontologies [45]		S7
		GeoVocab Ontology [46]		S40, S70
Analytic Concepts	Data Preprocessing & Integration			S14, S20, S27, S37, S41, S54, S80
	Data Analytics & Mining			S15, S17, S22, S24, S27, S35, S36, S37, S38, S41, S53, S54, S56, S60, S62, S63, S66, S67, S69, S74, S75, S76, S78, S79
	Control & Data Flow			S1, S3, S62, S75, S76, S78
Service Concepts	Software Component Management	Web Services	OWL-S Based [47]	S4, S11, S30, S31, S36
			WSDL-S Based [48]	S11
			SAWSDL based [49]	S20
			WSMO Based [50]	S4, S43
			Hydra Vocabulary Based [51]	S40
			Custom Models	S32, S37, S52
		Library Specific APIs		S27, S75, S79
		Generic APIs		S6
	Data Management	Multidimensional Data Schema		S10, S62, S71
		Graph Data Schema		S9, S46
	Composition Management	Workflow Templates		S4, S16, S20, S26, S53, S75, S78
		Provenance Related		S4, S25, S72
		Quality Related		S30, S31
	Implementation Management	Deployment Concepts		S3, S4, S8, S69
		Data Source		S3, S8, S10, S38, S43, S71
Intent Concepts	Analytic Query Expressions			S2, S9, S21, S48
	Analytics Requirements			S3, S7, S28, S44, S53, S68
	User Goals			S4, S39, S66, S75

conducting MapReduce[4] type of analytics. S62 proposes ontology derived from CRISP-DM terminology to represent control flow elements of a DAS.

4.2.3. Service Concepts

Service concepts capture knowledge related to DAS architectures and platforms. We identified 30 studies that model different aspects, namely web services, software APIs, data schemas, workflow design, knowledge related to provenance or data quality, deployment information and data sources. We classified service concept types into four subtypes which are software component management, data management, composition management and implementation management (see Table 2).

There are 10 studies under the *software component management* subtype, with the majority related to modeling web services that realize different operations in the analytics process. Many of those studies adapt or extend standard and popular semantic web service annotation standards: OWL-S [47], WSDL-S [48], SAWSDL [49], WSMO [50] and Hydra vocabulary [51]. Particularly, S31 extends OWL-S services in the area of event processing via the Complex Event Service Ontology. There are three studies [S32, S37, S57] that define custom concepts to represent web services used in DAS. S27 and S79 model Weka library specific software components. S6 provides the capability to model any generic API that can be used to implement DAS, through an ontology called Processing Element (PE) Knowledge Base. This ontology describes software components based on their input and output data types as well as relevant implementation details such as a URL for an HTTP request or a related JAVA class.

The second subtype represents *data management* concepts that support in representing and managing data structures and schemas. S10 proposes the BI Ontology to describe concepts such as dimension, hierarchy, level, property, measures that model data cubes related to Online Analytical Processing (OLAP) operations for a data warehouse. S62 proposes Corporate Data Model Ontology to capture metadata about data schemas useful in data access. S71 uses R2RML [5] to map relational databases into RDF data. Under the graph data schema category, S46 proposes the OpenCube ontology to describe concepts surrounding OLAP cubes in a data warehouse and similarly, S9 proposes to use an analytical schema to create RDF data warehouses.

The third subtype has concepts related to *composition management*, which model knowledge related to linking and executing multiple software services and data management operations together. Under that category, some studies propose different concepts (e.g.- Kepler ontology S4, Data Mining Workflow Ontology S78) to model scientific workflow templates and store their instances. Other composition management concepts are associated with provenance and quality. S4, S25 and S72 use ontologies to describe workflow provenance concepts. Quality related concepts are modeled in S30 and S31 to describe the quality and accuracy of different event-based services. We observed that S25 and S30 reuse concepts extended from the standard provenance ontology PROV-O [52].

The last subtype is related to the *implementation management* of an analytics solution. S3 uses a Deployment Ontology to describe the necessary deployment details for a MapReduce based system such as configuration variables, initial inputs, variables for profiling and performance measurements. S4 uses a Simulation Ontology to capture runtime metadata related to workflows. S69 models event processing framework components such as alert streams via an event ontology. S3, S10, S38, S43 and S71 propose concepts that describe the implementation of different data sources and how to access them. S8 proposes concepts to capture data source access details as well as system development details that represent the mappings between the database implementation and domain concepts.

4.2.4. Intent Concepts

Intent concepts capture knowledge with respect to the data analyst's requirements or goals. This knowledge can be in the form of low-level queries that need to be performed on data or high-level goals and intentions of users.

The first subtype represents concepts that relate to an *analytics query expression*. The Analytical Queries (AnQ) model in S9 facilitates expressing user queries that need to be performed on data. S21 and S48 propose to maintain a global ontology based on user-defined or standard concepts and use it to express user queries. S2 proposes the i2b2 Information Ontology, an intent related model that helps analysts to describe various dimensions of interest in data that are related to a particular task.

[4] https://research.google.com/archive/mapreduce.html
[5] https://www.w3.org/TR/r2rml/

Under *analytics requirements* subtype, S28, S44, S53 and S68 propose ontologies that capture user needs and constraints at a higher level. As an example, S44 uses an intent ontology called MIO (Multidimensional Integrated Ontology) which is auto-generated based on the topics, measures, and dimensions provided by the user. The KnowledgeDiscoveryTask class in Knowledge Discovery Ontology [S53] and the Problem component of the Decision Support Ontology [S68] are used as templates for instantiating analytics requirements. The Task-method Ontology proposed in S3 enables users to model desired methods that can be used to realize a particular task or define the expected role of a variable within a task. S7 uses the Measurement Ontology to capture concepts regarding product inspection and testing requirements based on ISO 9001 standards.

Under the *user goals* subtype, the Scientist's Intent Ontology in S4 and the Goal Oriented Model in S39, Purpose and Goal classes in DM3 Ontology [S66] and Goal component of the Base Ontology [S75] provide the capability to express a set of high-level user goals such as the desired outcomes of analytics tasks and the decision-making processes around them.

4.3. Sub-question 2: What tasks related to data analytics solution engineering are enabled using the identified concepts?

In this section, we analyze the association of 82 identified studies to different tasks related to data analytics solution engineering and how the semantic concepts discussed in section 4.2 are used to realize these tasks. The classification schema for analytics tasks was described in section 3.6, guided by the existing literature that defines data analytics and software engineering process.

Table 2 shows the mapping of 82 studies among 9 tasks and the different types of concepts. One study can be focused on more than one task, using one or more concept types. Business understanding is the focus of 6 studies that leverage domain or intent concepts. Data extraction and transformation approaches that use domain, analytics or service concepts are proposed in 15 studies. 31 papers propose data integration approaches, mostly using domain concepts, and some studies use analytics, service and intent concepts as well. Model selection (17 studies) and model building (15 studies) were conducted using domain, analytics or intent concepts. All four concept types were used to realize service composition (20 studies) and solution validation (8 studies). Code generation was supported by domain, analytics or service concepts in 4 studies. 9 studies that proposed approaches for result presentation and interpretation used one or two concept types out of four.

Different applications of those concepts are described in more detail in the rest of this section. The trend of publications related to each task is illustrated in Fig. 2. In 2014 and 2015 we can observe a special interest among researchers in applying semantic technology to support the engineering of DAS.

4.3.1. Domain Understanding

The domain understanding task focuses on analyzing the domain, context of the problem and understanding available datasets. This helps to establish solid definitions and facilitate the communication between different stakeholders. Moreover, the ontologies and concepts related to this task are inferable and the resulting knowledge has the flexibility of expanding over time.

Five methods [S5, S12, S18, S49, S59] use domain concepts for domain understanding. The platforms proposed in S5 and S59 use these concepts to capture semantic and interpretive aspects of data whereas S18 uses them to provide a standard specification of data for analysts. In contrast, S49 uses these concepts to model expert knowledge related to an analytics problem which is helpful in understanding the constraints and expected behavior. S12 proposes a feature-rich framework that can use custom-built domain concepts to understand the context thorough data browsing and visualization.

Two methods use intent concepts for domain understanding. S39 uses a Goal Model to define requirements that can help in understanding and designing a data warehouse model. S28 captures analytics requirements expressed in a natural language into an ontology which is refined through interviewing the stakeholders to identify data requirements for the analysis.

4.3.2. Data Extraction and Transformation

This task focuses on retrieving data from one or more sources and preparing it for the subsequent analysis. It includes transforming data into desired formats and annotating with additional metadata.

There are 8 studies that apply domain concepts for this task. S30, S33 and S70 annotate streaming input data using domain concepts making data queriable when necessary. In S18, data transformation is assisted by standard models built using domain concepts. In S2, S29 and S40, approaches for on-demand data extrac-

Table 2
Application of Concepts for Different Tasks Identified.

Related Task	Concept Classification			
	Domain Concept	Analytic Concept	Service Concept	Intent Concept
Domain Understanding	S5, S12, S18, S49	-	-	S28, S39
Data Extraction and Transformation	S2, S8, S18, S29, S30, S33, S34, S40, S50, S52, S59, S70, S80	S80	S8, S10, S38, S52	-
Data Integration	S2, S10, S12, S13, S19, S21, S23, S29, S34, S39, S40, S42, S43, S44, S45, S47, S48, S49, S51, S55, S59, S64, S65, S70, S76, S77, S81	S14	S9, S46, S71	S2, S9, S21, S39, S44, S48
Model Selection	S27, S49, S61, S62,	S15, S17, S22, S27, S37, S38, S41, S53, S56, S60, S62, S66, S75, S78, S79	-	S53, S66, S75
Model Building	S7, S58, S63, S64, S67, S69, S70, S73, S74, S77, S80, S82	S3, S54, S56, S63, S67, S69, S74, S80	-	S3, S7
Results Presentation and Interpretation	S35, S48, S49, S57, S61, S76	S35	S4, S25 ,S26	S4, S48
Service Composition	S11, S24, S27, S30, S31, S32, S43, S52, S62	S1, S20, S24, S27, S36, S41, S53, S62, S75, S78	S6, S11, S16, S20, S27, S30, S31, S32, S36, S37, S40, S43, S52, S53, S62, S75, S78	S53, S75
Analytics Solution Validation	S24, S61, S72, S76	S24, S75, S76	S25, S26, S72, S75	S4, S75
Code Generation	S47	S3	S3, S6, S38	-

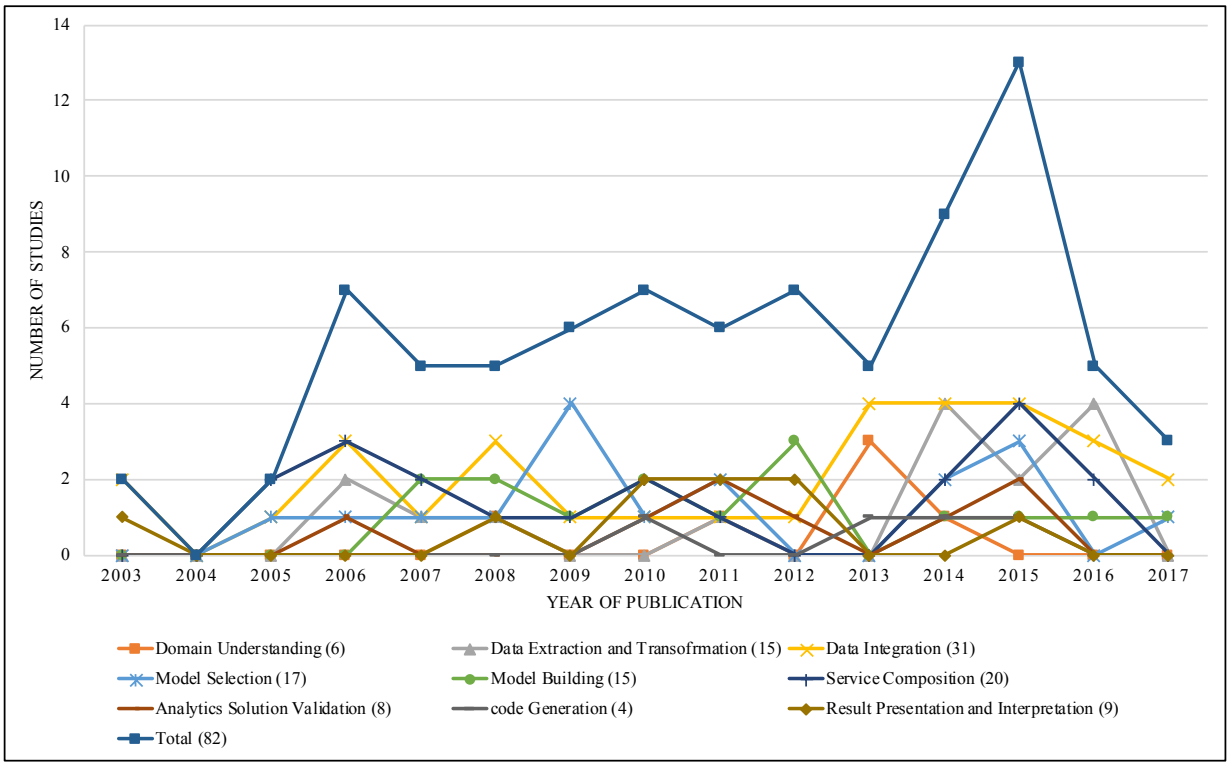

Fig. 2. Trend of Publications by Analytics Tasks

tion were proposed based on domain concepts that describe data sources. S34 focuses on the extraction of JSON data from web resources, generating semantic concepts around them and using those to convert data into ontology instances. S50 conduct data label correction using a domain ontology of medical entities. S59 enables analysts to understand datasets to help data extraction by querying knowledge represented in domain ontologies. S80 uses domain and analytics concepts together to access and transform static as well as streaming data.

Three studies use data source related service concepts for data extraction. In S10, these concepts are used to define the organization of data and how to access it on demand. S38 uses a set of concepts that can model any data source for model-driven data extraction code generation. S52 uses service concepts to identify implementations of required data transformations in a workflow.

S8 uses both domain and service concepts (data source and implementation management) for data extraction from heterogeneous sources such as event data streams and map the extracted data into a database schema. S52 uses domain concepts and service concepts related to data processing and to automatically generate new datasets on demand.

4.3.3. Data Integration

Data integration implies combining heterogeneous datasets in order to obtain a high-level and coherent view of the data. Most studies use domain concepts to aid in this task. Some studies use intent concepts to conduct integration based on user needs. S9, S14, S46 and S71 are special cases that leverage analytics and service concepts to support the integration task.

We identified five different strategies in which intermixed concepts from different classes contribute to data integration. The first one, observed in S12, S13, S19, S39, S42, S46 and S47, transforms data from heterogeneous sources into instances of a global ontology. S12, S13, S39 and S42 conduct ETL processes incorporating this global ontologies to create semantic aware data warehouses.

The second one, identified in S2, S21, S23, S44, S48 and S77 uses a global ontology that represents the user's perspective and a set of local ontologies to represent datasets. Then by aligning or converting local ontologies into the global ontologies, the data is matched accordingly. The users can refer to the global ontology to query the data.

The third approach describes each dataset through a local ontology and achieves data integration by merging local ontologies together. S34 is an example where each local ontology is constructed by first extracting data provided in JSON format, generating suitable semantic concepts and using those concepts to convert the extracted data into ontology instances.

The fourth method used in 10 studies, S10, S14, S29, S40, S45, S51, S59, S64, S70, S76, maintains linked meta-data about datasets from different data sources so that relevant data can be acquired at query time from multiple sources.

The fifth one in S9, S43, S49, S55, S65, S71 and S81 is a query or requirement driven approach for data integration where formal rules or program logic are used to represent user queries and analytics questions. Different datasets are mapped into those rules to derive answers. S9 is unique in that it captures analytics queries as intent concepts.

4.3.4. Model Selection

Model selection is crucial for non-expert users who do not have intuitive knowledge about the performance of different models in different contexts or when there are many competing analytics models and techniques that can be used for a single purpose. This task facilitates comparison of algorithms or makes recommendations of tools and models suitable for users.

There are 4 studies that use domain concepts for this purpose. S49 stores expert domain knowledge in a domain ontology and uses that to evaluate possible analytics models generated by associate rule mining, incorporating an "interestingness" measure. S61 proposes a recommendation engine that maintains a repository meta-data related to historical analytic solutions using domain concepts and when a user provides a new dataset, a matching solution is recommended to the user. In S27, suitable analytics technique for a particular dataset is identified by matching dataset features represented in domain concepts with the requirements of the analytics techniques captured through analytics concepts. S62 supports model selection by allowing users to define the context of analytics problem via domain concept and using analytics concepts to recommend appropriate analytics techniques.

There are 15 methods that apply data analytics and mining focused concepts for model selection. The simplest method proposed in S17 uses an ontology to describe data analytics algorithms and creates a knowledge repository to provide a querying capability for the users.

Concepts defined in S15, S27, S37, S41, S53, S60, S66, S75 and S78 assist users in matching analytics components that suit their goal and constraints. S53, S66 and S75 are unique among those studies as they propose intent concepts to capture user goals and requirements which are then matched with suitable models represented as analytics concepts.

In S22, analytics concepts are used to model data mining algorithms which are then linked to web services, providing the means for service composition.

S38 and S79 use analytics concepts to capture existing analytics workflows so novice users can search and learn from them.

4.3.5. Model Building

Model building represents a core task during analytics process development, where the selected model needs to be applied and customized for the problem at hand.

S58, S64, S70, S73, S77 and S82 use domain concepts to write analytic queries or event patterns executed on data to get descriptive analytics insights. S67, and S69 follows a similar approach, but use analytics concepts to support users in writing queries or rules.

S63 and S74 use domain concepts and analytics concepts to describe and customize feature space for analytics model construction.

S7 leverages domain concepts and intent concepts for model building. The model generated consists of axioms based on formal competency questions to evaluate whether an enterprise model represented through domain concepts adhere to the intent concepts that align with different compliance standards.

S56 uses analytics concepts to define analytics experiments in detail for supervised classification of propositional datasets. S54 proposes a case based system for model selection and uses analytics concepts to adapt the solutions suggested by case based reasoning to fit user's interest. S3 applies analytics concepts (together with intent concepts) for analytics model generation. The purpose is to facilitate the modeling of a MapReduce based analytics solution.

4.3.6. Result Presentation and Interpretation

Having semantic models that store knowledge related to different aspects of the analytics process inherently provides a certain inference and interpretation capability that helps in result representation. Here we are looking at studies that specify particular methods that use semantic concepts to explicitly facilitate result presentation and interpretation.

Six studies use domain concepts for result presentation and interpretation. S49 uses domain concepts to store expert knowledge which is then used for validation by aligning it with the data mining results and evaluate their interestingness. S57 uses domain concepts to interpret hypotheses and related attributes on statistical datasets. S61 use domain concepts to annotate data tables via ontology alignment, enabling easy interpretation. S35 incorporate domain concepts with analytics concepts to generate reports on the conducted data analytics tasks. S76 generates metadata about numerical analysis using domain and analytics concepts. S48 uses domain concepts with intent concepts to extract meta-data about OLAP operations and generate reports. Further, S48 proposes a method to automatically match the OLAP report with other documents in a related repository.

When we look at the use of service concepts for presenting results, S4, S25 and S26 capture the knowledge on different aspects of scientific workflows, especially those that can help to describe and present the outputs/results.

In S4 intent concepts are used to annotate workflows with initial goals of the analysts, in order to identify different decisions that have led to the outcome and explain the results from the perspective of the analyst.

4.3.7. Service Composition

Service composition means identifying and putting together different analytics related services to provide a complete or partial DAS- from data acquisition and extraction to results generation. Scientific workflow planning and service composition are largely incorporated into this task.

Identifying a suitable service or tool to include in a DAS is a major activity within service composition. Some studies [S32, S36, S37, S40] use software component management concepts to guide component selection, but leave the responsibility of process composition to the analyst. S32 proposes a model to represent and recommend web services using predefined rules based on a context expressed through domain concepts. S37 and S40 use service concepts to model a wide array of software components to be selected from, including pre-processing capabilities such as null value removal. Users can query the ontologies to identify suitable components. S36 proposes a methodology to facilitate the selection of suitable service implementations based on the input data. In S43 and S52, suitable service implementations for datasets

are identified by matching characteristics represented through domain and service concepts.

In contrast, S30 and S31 follow an approach which facilitates the selection of data sources that match the existing software components by the ability to extract and process related data. Data provision services are annotated using OWL-S and SSN ontology concepts so that users can query them and identify suitable services. Moreover S30 incorporates quality related service concepts to represent important quality attributes that need to be considered in component selection.

Other studies extend service selection to incorporate different domain, service and analytics concepts and facilitate composition planning and execution.

S24 uses domain concepts that describe datasets, data analytics and data mining concepts to support workflow composition through matching the analytics operations with data properties, on top of the Kepler workflow composition canvas. S27 uses domain, analytics and service concepts to select suitable data sources, analytics techniques and software components respectively. S41 uses analytics concepts for representing components of the Weka analytics tool and linking these components as a process. It does not provide executable workflows but recommends an analytics plan to be manually executed by the user. S1 proposes a similar approach for the MIT Lincoln Laboratory's Composable Analytic Environment that including executable workflow definitions. S20 uses service concepts for software component modeling together with analytics concepts for modeling data transformation rules between software components. S62 users Control and Data flow concepts to guide the knowledge discovery process and supports decision making at each stage using a knowledge base that encompasses domain, analytic, and service concepts.

S78 uses analytics and workflow template concepts to generate analytics processes, with a major focus on performance optimization. S16 uses concepts related to workflow templates to store pre-composed analytics processes which can be queried by users in order to select a suitable implementation.

S53 captures analytics, service and intent concepts via a Knowledge Discovery Ontology and offers support for planning an abstract analytics process. S6 supports software composition through components modeled as generic APIs by matching respective inputs and outputs. It assists users in planning and matching analytics components with a comprehensive goal based planning method but this is limited by the inability to incorporate different parameters other than the input and output conditions. Similarly, S75 uses analytic, service and intent concepts to capture the user goals and KDD workflows implemented in RapidMiner. These concepts are used to identify optimal analytics process for new analytics problems based on Hierarchical Task Network planning.

S11 provides comprehensive scientific workflow composition facilities including a graphical user interface but depends on the Taverna workflow engine, SADI/BioMoby plug-ins and web services that are SADI-compliant. It includes web service concepts to model components that implement data analytics algorithms and domain concepts to define input and output data. Those concepts are used to recommend services that match analytics requirements or data constraints.

4.3.8. Analytics Solution Validation

Validation of the analytics solutions involves capturing provenance data, validating solution workflows for service compatibility or data consistency and confirming the solution addresses the analyst's goals. S61 uses domain ontology to generate metadata about input data and output results, enabling provenance. S24 uses domain concepts that describe data and analytics concepts to validate the structural and semantic correctness of a workflow before execution. S76 capture provenance data related to datasets, data sources as well as operations around numerical analysis through domain and analytics concepts. S25 uses provenance-related concepts to model workflow and data, store them as RDF triples to allow users to query them in order to validate workflows, identify defects or extract further information. S26 defines a Research Object as an instance of a scientific workflow, for provenance purposes. S75 proposes solution validation approach as an extension for RapidMiner tool that use analytic, service and intent concepts to annotate data, operators, models, data mining tasks and KDD workflows. The Scientist's Intent Ontology in S4 uses goal focused intent concepts to describe user goals that are used for validation of workflows. S72 uses domain and service concepts to capture provenance of ETL workflows.

4.3.9. Code Generation

Methods that support code generation rely on ontologies to convert abstract models into the executable analytics software. This is an important task as it reduces the burden of software programming for data analysts. Code generation can be used to support one or multiple stages of an analytics process (workflow) execution.

Most techniques use service concepts (e.g. S3, S6 and S38) to drive code generation in a Model Driven Engineering (MDE) fashion. For example, data sources modeled as service concepts in S38 are used to generate data extraction software modules. In S6 generic API modeling service concepts are used to generate an executable analytics process. In S3, analytics as well as service concepts (deployment and data source related) capture the implementation details related to each analytics task. This enables a semi-automated code generation scheme for selected analytics techniques. The method proposed in S47 is the only one that uses domain concepts to model data sources, which are then used for generating code for data extraction from data sources into linked datasets.

5. Discussion

5.1. Limitations of Existing Work

This section summarizes the limitations we identified by studying what type of semantic concepts were used in data analytics solutions and how those different concepts types were applied in different analytics tasks.

5.1.1. Limited Usage of Intent Concepts

Although there are 14 studies that propose intent concepts, (see Table 2), we can observe that only a few studies are using intent concepts at different stages of the analytics process, with the exception of data integration. The survey did not identify any studies that use intent concepts for data extraction and transformation, although this is a computationally expensive and time consuming task that may waste resources if not performed aptly. Existing techniques use intent concepts mostly on facilitating the search of algorithms, data providers, web services, and computational software modules. Hence to a large extent, analytics requirements such as what business decisions will be supported by this analysis or what level of accuracy is required, are still a part of the mental model of the developer or the analyst who performs these tasks. In practice, several iterations of data cleansing, reformatting, model selecting and process composition may be required in order to optimally address the analytics problem at hand. This may result in less effective DASs whose performance is likely to degrade with time. In addition, modifying the process can only be conducted by someone with a sound understanding of the original analytics requirements. Moreover, as discussed in [53] cognitive and context information, which can be captured through intent concepts, is crucial for accurate interpretation and validation of data mining knowledge. We believe that incorporating suitable intent concepts further can enhance the efficiency and effectiveness of the DAS engineering.

5.1.2. Lack of Proper Concept Classification

Semantic concepts can be classified in many different ways. For an example, S43 separates domain, analytics and service knowledge in three ontologies and S79 uses a class hierarchy to separate data mining related entities as process, information content and realizable entities. Yet in some studies, this separation of concept types is not clearly visible. One ontology with a unique prefix may contain concepts related to one or more categories without attempting to follow modular approaches such as class hierarchies. For example, S37 model both analytics and service concepts in one ontology, and S3 model both analytics and intent concepts in the Task-methods ontology. S27 contains two ontologies (WekaOntology and ProtOntology) that cut across concepts of all classes without proper separation.

5.1.3. Little Support for End-to-End Development Process

Though there is an array of research on adapting semantic models for different development tasks such as data integration or model selection, only a few studies seem to go beyond addressing one or two tasks in the development lifecycle. In many cases, knowledge from previous tasks would have been very useful if carried over to the next tasks. Hence there is a lack of studies that propose semantic modeling based solutions to support the DAS engineering lifecycle.

We identified 4 studies that use semantic models for code generation (section 4.4.9), related to data transformation and analytics process execution. They are also limited to a specific domain or a tool and do not provide sufficient flexibility to be used for a wider class of DASs.

5.2. Recommendations for Future Research

Upon the findings of this study, we propose a set of recommendations for future research regarding the application of semantic models for DAS engineering.

5.2.1. Developing Intent Concepts for Analytics

As discussed in section 5.1.1, service composition techniques among the identified studies do not leverage intent concepts adequately. One reason could be that intent concepts are of too high-level (e.g. a business goal) or low level (a query). As the data analytics community is extending wider into different industries and organizations and as analytics contexts and requirements are changing rapidly, it is necessary to explore techniques that consider all dimensions such as business requirements, contexts and constraints. Hence a potential research area is to study how high level user goals and context can be represented and incorporated in DAS engineering through data integration, process construction, and result interpretation. Initial work in this direction can be found at Bandara et. al. [54].

Such approaches of linking the user intentions and contexts into analytics models have the potential of changing static analytics models deployed today into dynamic and adaptable analytics models, responding to changes in user goals or the operational context.

5.2.2. Decoupling Concept Classes and Encouraging Concept Reuse Across Development Tasks

In section 5.1.2 we discussed that it is more effective to decouple different concepts as separate ontologies can lead to better and modular knowledge management. As a result, each concept type can be reused or evolve independently of the others, enabling users to change the application domain, implementation, data source or the analytics requirements without altering other models. Then the integration between those different knowledge areas has to be done separately within the DAS environment, considering the context as well. Some studies achieve concept integration through program logic or annotation schemes, but it would be useful to have standard, platform-independent ways of modeling the relationships between different types of analytics knowledge to match the context of a particular analytics process.

To promote the usage of semantic models among the research community and to enhance the value and the reusability of the research, it is essential to promote the reuse of ontologies. This enables the creation of a common vocabulary and the resulting data/models become interoperable among a variety of systems. We observed certain ontologies like SSN [42] and Gene Ontology[43] are being used in multiple research studies. There are ongoing efforts such as OBO Foundry [6] that recommends re-using classes already defined in other ontologies classes. Yet we believe there is room for the system development community to adapt more concepts from well-developed ontologies, particularly analytics concepts proposed in ontologies such as OntoKDD S60 and OntoDM S79, to improve user support for analytics process design.

As knowledge represented through ontologies can enhance each task of the DAS engineering process, a standard framework for designing and extending ontologies that are usable in all analytics process stages is necessary. The ontologies should incorporate knowledge related to domain concepts and business goals as well as the concepts useful for the execution level. For example, an ontological representation of a data source may contain information necessary to retrieve data, but also information about the data quality, the latency of data acquisition, metadata that can be used to decide which algorithm is suitable to process the data (e.g.the knowledge of whether the data is time-series or not can reduce the space of algorithms we can use to process it) and the relationship between the data and other concepts. Representation of existing knowledge and enabling efficient reuse of accumulated knowledge and resources can reduce the cost for an organization, otherwise spent on expert consultations or employee training.

5.2.3. Semantic Model-Driven Data Analytic Solution Engineering

Finally, our evidence reveals the opportunities of using semantic models for code generation in the light of model-driven engineering methods. This needs to be explored and experimented further as it has the potential of lifting the burden of software programming expertise from data analysts. It can lead to a significant cost reduction and resource utilization related to software development efforts in analytics solution engineering. There are already some examples of applying MDE for data analytics applications [55], such as creating Hadoop MapReduce analysis through conceptual models [56]. A promising finding is that four studies that are identified related to code generation (section 4.3.9) use semantic models that are well aligned with the four ontologies proposed by Pan et.al, in their book Ontology-Driven Software Engineering [17]. They align Requirement ontology (intent concepts) to Computational Independent Model (CIM), Infrastructure Ontology (service concepts) to Platform

[6] www.obofoundry.org/

Specific Model (PSM) and propose to use a domain ontology for converting CIM to Platform Independent Model (PIM) and a business process ontology (analytics concepts) to convert PIM to PSM. Hence we believe that studying the use of ontologies to develop analytics solutions in a model driven fashion, particularly adapting the framework proposed by Pan et.al [17], is timely and significant.

Semantic based service orchestration plays a significant role in realizing a semantic model driven analytics environment, as all operations from data exporting, integration into model building, execution and result publication can be done as independent services modules represented through semantic models. Yet there is a lack of applications that integrate the existing body of research related to semantic based service orchestration such as [57–59] with semantic data analytics process construction research. Such a combination can contribute to a paradigm of service-based data analytics solutions and establish the basis for semantic model driven data analytics systems.

6. Limitations of the Study

There are some limitations of our study, mainly due to the literature selection process, including the selection of keywords and the construction of inclusion and exclusion criteria. Firstly, our study focuses on peer-reviewed publications in academic literature only. So gray literature such as technical reports, white papers and unpublished work were not included. Secondly, the study might be missing some relevant work due to the search string failing to match other relevant within the digital libraries. A snowballing technique has helped to eliminate this limitation to a certain level. These limitations are in-line with our exclusion criteria, yet they pose a risk for the completeness and validity of the results.

One other limitation is that the selected databases may not contain all related literature, especially applications published in domain specific venues such as medical journals. We attempted to reduce the impact of such limitations by using the of web of science database, which exposes our search query into diverse disciplines. We are coming from software engineering and modeling background and hence the paper selection and mapping may be biased to that point of view in spite of the efforts made to conduct an unbiased literature filtering process, especially in the cases where the inspected papers do not provide a clear-cut definition of their research problem and proposed approach.

As we conduct a systematic process to identify and map literature, this paper may contain some outdated work or not reflect the most recent achievements in the discipline. Recent ontologies published related to analytics and data modeling such as OntoDT [60] were not observed in any identified work. This may be due to the limitation of search approach and authors believe there will be more future research that utilizes existing analytics related knowledge in analytics solution engineering.

As the goal of this study is to present a holistic overview on how semantic modeling has been used in engineering analytics solutions, summarizing two decades of research, it is beyond the scope of this paper to drill down certain specific characteristics of analytics solutions that support particular tasks and conduct a thorough evaluation. We limit our contribution to a mapping study which can be used by researchers to study certain aspects extensively in the future.

7. Conclusion

Capturing knowledge using models to drive the software development life cycle is at the heart of the software engineering discipline. Traditional models have serious limitations in the area of building data analytics solutions, which are characterized by the need to represent rich knowledge encompassing specialized application domains, complex computing infrastructures and changing analytics requirements. This has triggered our interest in the use of semantic modelling and ontologies as a way of underpinning new software development practices in this area. In this paper, we presented 82 studies identified through a systematic mapping study, that leverage semantic modeling for engineering data analytics solutions. We adopted a broad approach encompassing distinct research areas such as data mining and service computing.

The results of our study reveal the diversity of knowledge representation in existing studies. Through sub-question 1 we identified what type of semantic concepts are modelled and used in the literature. They were falling under four main categories: domain, analytic, service and intent ontology. Through sub-question 2 we identified the different categories of analytics or software engineering related tasks mentioned by the identified literature. Different types of concepts were observed to play different roles in improving each

task and supporting various stages of the DAS engineering. Semantic modeling was highly used for tasks such as data integration, model selection, process composition and data extraction, which shows the ability of semantic models to represent heterogeneous resources. Studies that focus on model selection and process composition tasks highlight the capability of semantic models to provide end user support for analytics solution engineering. We identified and discussed some limitations in existing work such as the limited usage of intent concepts and the lack of end-to-end support for analytics process engineering.

Recommended future work, discussed in section 5.2, emphasizes the importance of moving semantic technology out of certain research silos and aiming at developing new research agendas around capturing high-level intents and goals of data analysts and translating them to executable analytics processes, incorporating a multitude of well-defined semantic knowledge repositories that have the capacity to be developed, expanded and maintained independently from each other. This can be achieved within established software engineering frameworks that will need to be specifically tailored to the particular characteristics of the DAS engineering life-cycle as presented in [61]. As the next stage, we are working on designing a requirement driven platform that provides support for end-to-end analytics process engineering, incorporating semantic concept types identified through this mapping study [54, 59].

8. List of Included Studies

S1 K. K. Nam, R. T. Locke, S. Yenson, K. Chang, and L. H. Fiedler, "Advisory services for user composition tools," In Semantic Computing (ICSC), 2015 IEEE International Conference on, pp. 9-16, 2015.

S2 J. G. Klann, A. Abend, V. A. Raghavan, K. D. Mandl, and S. N. Murphy, "Data interchange using i2b2" Journal of the American Medical Informatics Association, vol. 23, pp. 909-15, Sep 2016.

S3 T. W. Wlodarczyk, C. Rong, B. Jia, L. Cocanu, C. I. Nyulas, and M. A. Musen, "DataStorm An Ontology-Driven Framework for Cloud-Based Data Analytic Systems," 2010 6th World Congress on Services (SERVICES-1), pp. 123-127, 2010.

S4 E. Pignotti, P. Edwards, N. Gotts, and G. Polhill, "Enhancing workflow with a semantic description of scientific intent," Web Semantics: Scinece, Services and Agents on the World Wide Web, vol. 9, pp. 222-244, 2011.

S5 E. W. Kuiler, "From big data to knowledge: an ontological approach to big data analytics," Review of Policy Research, vol. 3, pp. 311-318, 2014.

S6 P. Coetzee and J. Stephen, "Goal-based analytics composition for on-and off-line execution at scale," Trustcom/BigDataSE/ISPA, 2015 IEEE, vol. 2, pp. 56-65, 2015.

S7 H. M. Kim, M. S. Fox, and A. Sengupta, "How to build enterprise data models to achieve compliance to standards or regulatory requirements (and share data)," Journal of the Association for Information Systems vol. 8, p. 105, 2007.

S8 C. Pautasso and O. Zimmermann, "Linked Enterprise Data Model and its use in Real Time Analytics and Context-Driven Data Discovery," IEEE Software, vol. 32, pp. 7-9, 2015.

S9 D. Colazzo, F. Goasdoue, I. Manolescu, and A. Roatis, "RDF Analytics: Lenses over Semantic Graph," Proceedings of the 23rd international conference on World wide web, pp. 467-478, 2014.

S10 D. Sell, D. C. d. Silva, F. D. Beppler, M. Napoli, F. B. Ghisi, R. Pacheco, et al., "SBI: a semantic framework to support business intelligence," In Proceedings of the first international workshop on Ontology-supported business intelligence, p. 11, 2008.

S11 D. Withers, E. Kawas, L. McCarthy, B. Vandervalk, and M. Wilkinson, "Semantically-guided workflow construction in Taverna: the SADI and BioMoby plug-ins," Proceedings of the 4th international conference on Leveraging applications of formal methods, verification, and validation, vol. 6415, pp. 301-312, 2010.

S12 S. Andrews, "The CUBIST Project: Combining and Uniting Business Intelligence with Semantic Technologies," International Journal of Intelligent Information Technologies, vol. 9, pp. 1-15, 2013.

S13 R. P. Deb Nath, K. Hose, and T. B. Pedersen, "Towards a Programmable Semantic Extract-Transform-Load Framework for Semantic Data Warehouses," Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP pp. 15-24, 2015.

S14 B.L. Do, T.D. Trinh, P. R. Aryan, P. Wetz, E. Kiesling, and A. M. Tjoa, "Toward a statistical data integration environment," Proceedings of the 11th International Conference on Semantic Systems, pp. 25-32, 2015.

S15 M. V. Nural, M. E. Cotterell, and J. A. Miller, "Using Semantics in Predictive Big Data Analytics.," In Big Data (BigData Congress), 2015 IEEE International Congress, pp. 254-261, 2015.

S16 Y. Gil, V. Ratnakar, E. Deelman, M. Spraragen, and J. Kim, "Wings for Pegasus: A Semantic Approach to Creating Very Large Scientific Workflows," OWL: Experiences and Directions (OWLED), 2006.

S17 M.S. Lin, H. Zhang, and Z.G. Yu., "An Ontology for Supporting Data Mining Process," Computational Engineering in Systems Applications, IMACS Multiconference on, vol. 2, pp. 2074–2077, 2006.

S18 R. G. Wang, W. D. Dai, and J. R. Cheng, "An Ontology-Based Data Mining Framework in Traffic Domain," Applied Mechanics and Materials, vol. 121-126, pp. 55-59, 2011.

S19 A. Bouchra, A. A. Wakrime, A. Sekkaki, and K. Larbi, "Automating Data warehouse design using ontology," International Conference on Electrical and Information Technologies, 2016.

S20 S. Shumilov, Y. Leng, M. El-Gayyar, and A. B. Cremers, "Distributed Scientific Workflow Management for Data-Intensive Applications," 12th IEEE International Workshop on Future Trends of Distributed Computing Systems, FTDCS'08, vol. 12, pp. 65-73, 2008.

S21 J. A. R. Castillo, A. Silvescu, D. Caragea, J. Pathak, and V. G. Honavar, "Information extraction and integration from heterogeneous, distributed, autonomous information sources - a federated ontology-driven query-centric approach," IEEE International Conference on Information Reuse and Integration, pp. 183-191, 2003.

S22 X. Gong, T. Zhang, F. Zhao, L. Dong, and H. Yu, "On Service Discovery for Online Data Mining Trails," Second International Workshop on service discovery for online data mining trails, vol. 1, pp. 478-482, 2009.

S23 M. Gagnon, "Ontology-based integration of data sources," 10th International Conference On Information Fusion, pp. 1-8, 2007.

S24 J. Zhang, "Ontology-Driven Composition and Validation of Scientific Grid Workflows in Kepler-a Case Study of Hyperspectral Image Processing," Fifth International Conference on Grid and Cooperative Computing Workshops (GCCW'06.), 2006.

S25 A. Chebotko, S. Lu, X. Fei, and F. Fotouhi, "RDFProv: A relational RDF store for querying and managing scientific workflow provenance," Data & Knowledge Engineering, vol. 69, pp. 836-865, 2010.

S26 K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, et al., "Using a suite of ontologies for preserving workflow-centric research objects," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 32, pp. 16-42, 2015.

S27 M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, and P. Veltri, "Using ontologies for preprocessing and mining spectra data on the Grid," Future Generation Computer Systems, vol. 23, pp. 55-60, 2007.

S28 N. Kushiro, "A Method for Generating Ontologies in Requirements Domain for Searching Data Sets in Marketplace," 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW), pp. 688-693, 2013.

S29 P. Gong, W. Qu, and D. Feng, "An Ontology for the Integration of Multiple Genetic Disorder Data Sources," Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, pp. 2824–2827, 2006.

S30 D. Puiu, P. Barnaghi, R. Tonjes, D. Kumper, M. I. Ali, A. Mileo, et al., "CityPulse: Large Scale Data Analytics Framework for Smart Cities," IEEE Access, vol. 4, pp. 1086-1108, 2016.

S31 F. Gao, M. I. Ali, and A. Mileo, "Semantic discovery and integration of urban data streams," Proceedings of the Fifth International Conference on Semantics for Smarter Cities, vol. 1280, pp. 15-30, 2014.

S32 Z. Laliwala, V. Sorathia, S. Chaudhary, and t. I. I. C. o. Distributed, "Semantic and Rule Based Event-driven Services-Oriented Agricultural Recommendation System," 26th IEEE International Conference on Distributed Computing Systems Workshops, 2006, pp. 24-24, 2006.

S33 S. Qanbari, N. Behinaein, R. Rahimzadeh, and S. Dustdar, "Gatica: Linked Sensed Data Enrichment and Analytics Middleware for IoT Gateways," 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), pp. 38-43, 2015..

S34 Y. Yao, H. Liu, J. Yi, H. Chen, X. Zhao, and X. Ma, "An automatic semantic extraction method for web data interchange," 6th International Conference on Computer Science and Information Technology (CSIT), pp. 148-152, 2014.

S35 T. Kliegr, V. Svatek, M. Ralbovsky and M. Simunek, "SEWEBAR-CMS: semantic analytical report authoring for data mining results", Journal of Intelligent Information Systems, vol. 37, no. 3, pp. 371-395, 2010.

S36 T. Marinho, E. B. Costa, D. Dermeval, R. Ferreira, L. M. Braz, I. I. Bittencourt, et al., "An ontology-based software framework to provide educational data mining," in 2010 ACM Symposium on Applied Computing, 2010, pp. 1433-1437.

S37 B. T. G. S. Kumara, I. Paik, J. Zhang, T. H. A. S. Siriweera, and K. R. C. Koswatte, "Ontology-Based Workflow Generation for Intelligent Big Data Analytics," pp. 495-502, 2015.

S38 J.N. Mazon, J. Jacobo Zubcoff, L. Garriga, and R. Espinosa, "Knowledge Spring Process - Towards Discovering and Reusing Knowledge within Linked Open Data Foundations," pp. 291-296, 2014.

S39 L. Bellatreche, K. Selma and Nabila Berkani. "Semantic data warehouse design: From ETL to deployment a la carte.", International Conference on Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2013.

S40 . M. Frank and S. Zander, "Smart Web Services for Big Spatio-Temporal Data in Geographical Information Systems," ESWC workshop: Services and Applications over Linked APIs and Data (SALAD 2016), 2016.

S41 F. P. Abraham Bernstein, Shawndra Hill. , "Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification ," IEEE Transactions on knowledge and data engineering, vol. 17, pp. 503-518, 2005.

S42 S. K. Bansal, "Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration," 2014 Ieee International Congress on Big Data (Bigdata Congress), pp. 521-528, 2014.

S43 D. Sell, L. Cabral, E. Motta, J. Domingue and R. Pacheco, "Adding semantics to business intelligence.," Proceedings. Sixteenth International Workshop on Database and Expert Systems Applications,, pp. 543-547, 2005.

S44 V. Nebot, R. Berlanga, J. M. Perez, M. J. Aramburu and T. B. Pedersen, "Multidimensional integrated ontologies: A framework for designing semantic data warehouses", In Journal on Data Semantics XIII, pp. 1-36, 2009.

S45 D. Skoutas and A. Simitsis. "Designing ETL processes using semantic web technologies", Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. ACM, 2006.

S46 L. Etcheverry and A. Vaisman. "Enhancing OLAP analysis with web cubes." The Semantic Web: Research and Applications, pp. 469-483, 2012.

S47 A. Harth, C. A. Knoblock, S. Stadtmuller, R. Studer and P. Szekely, "On-the-fly integration of static and dynamic linked data. ," In Proceedings of the Fourth International Conference on Consuming Linked Data, vol. 1034, pp. 1-12, 2013.

S48 T. Priebe and G. Pernul. "Ontology-based integration of OLAP and information retrieval." Proceedings of 14th International Workshop on-Database and Expert Systems Applications, IEEE, 2003.

S49 L. Brisson and M. Collard, "An ontology driven data mining process," In International Conference on Enterprise Information Systems, pp. 54-61, 2008.

S50 M. Guo, Y. Liu, J. Li, H. Li, and B. Xu. "A knowledge based approach for tackling mislabeled multi-class big social data." In European Semantic Web Conference, pp. 349-363. Springer, Cham, 2014.

S51 S. K. Ramnandan, A. Mittal, C. A. Knoblock, and P. Szekely. "Assigning semantic labels to data sources." In European Semantic Web Conference, pp. 403-417. Springer, Cham, 2015.

S52 J. L. Ambite and D. Kapoor. "Automatically composing data workflows with relational descriptions and shim services." In The Semantic Web, pp. 15-29. Springer, Berlin, Heidelberg, 2007.

S53 M. Zakova, Monika, P. Kremen, F. Zelezny, and N. Lavrac. "Automating knowledge discovery workflow composition through ontology-based planning." IEEE Transactions on Automation Science and Engineering Vol. 8(2),pp. 253-264, 2011.

S54 M. Charest, S. Delisle, O. Cervantes, and Y. Shen. "Bridging the gap between data mining and decision support: A case-based reasoning and ontology approach." Intelligent Data Analysis Vol. 12(2), pp. 211-236, 2008.

S55 J. P. Calbimonte, O. Corcho, and A. J. Gray. "Enabling ontology-based access to streaming data sources." In International Semantic Web Confer-

ence, pp. 96-111. Springer, Berlin, Heidelberg, 2010.

S56 J. Vanschoren, and L. Soldatova. "Expose: An ontology for data mining experiments." In International workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010), pp. 31-46. 2010.

S57 H. Paulheim, "Generating possible interpretations for statistics from linked open data." In Extended Semantic Web Conference, pp. 560-574. Springer, Berlin, Heidelberg, 2012.

S58 Q. Zhou, Y. Simmhan, and V. Prasanna. "Incorporating semantic knowledge into dynamic data processing for smart power grids." In International Semantic Web Conference, pp. 257-273. Springer, Berlin, Heidelberg, 2012.

S59 D. A. Dimitrov, J. Heflin, A. Qasem, and N. Wang. "Information integration via an end-to-end distributed semantic web system." In International Semantic Web Conference, pp. 764-777. Springer, Berlin, Heidelberg, 2006

S60 C. Diamantini, D. Potena, and E. Storti. "Kddonto: An ontology for discovery and composition of kdd algorithms." Third Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD'09) ,pp. 13-24, 2009.

S61 Y. Gil, P. Szekely, S. Villamizar, T. C. Harmon, V. Ratnakar, S. Gupta, M. Muslea, F. Silva, and C. A. Knoblock. "Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows." In International Semantic Web Conference, pp. 65-80. Springer, Berlin, Heidelberg, 2011.

S62 M. Choinski, and J. A. Chudziak. "Ontological learning assistant for knowledge discovery and data mining." In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on, pp. 147-155. IEEE, 2009.

S63 A. Tsymbal, S. Zillner, and M. Huber. "OntologyâĂŞSupported Machine Learning and Decision Support in Biomedicine." In International Conference on Data Integration in the Life Sciences, pp. 156-171. Springer, Berlin, Heidelberg, 2007.

S64 A. Albalushi, R. Khan, K. McLaughlin, and S. Sezer. "Ontology-based approach for malicious behaviour detection in synchrophasor networks." In Power & Energy Society General Meeting, 2017 IEEE, pp. 1-5. IEEE, 2017.

S65 M. Rodriguez-Muro, R. Kontchakov, and M. Zakharyaschev. "Ontology-based data access: Ontop of databases." In International Semantic Web Conference, pp. 558-573. Springer, Berlin, Heidelberg, 2013.

S66 Y. Li, M. A. Thomas, and K. M. Osei-Bryson. "Ontology-based data mining model management for self-service knowledge discovery." Information Systems Frontiers, Vol. 19(4), pp. 925-943, 2017.

S67 J. Debattista, S. Scerri, I. Rivera, and S. Handschuh. "Ontology-based Rules for Recommender Systems." In SeRSy, pp. 49-60. 2012.

S68 M. Rospocher and L. Serafini. "Ontology-centric Decision Support." In SeRSy, pp. 61-72. 2012.

S69 K. Taylor and L. Leidinger. "Ontology-driven complex event processing in heterogeneous sensor networks." In Extended Semantic Web Conference, pp. 285-299. Springer, Berlin, Heidelberg, 2011.

S70 F. Lecue, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi, and M. Sbodio. "Predicting severity of road traffic congestion using semantic web technologies." In European Semantic Web Conference, pp. 611-627. Springer, Cham, 2014.

S71 Y. Khan, A. Zimmermann, A. Jha, D. Rebholz-Schuhmann, and R. Sahay. "Querying web polystores." In Big Data (Big Data), 2017 IEEE International Conference on, pp. 3190-3195. IEEE, 2017.

S72 A. Freitas, B. Kampgen, J. G. Oliveira, S. O'Riain, and E. Curry. "Representing interoperable provenance descriptions for ETL workflows." In Extended Semantic Web Conference, pp. 43-57. Springer, Berlin, Heidelberg, 2012.

S73 K. Teymourian and A. Paschke. "Semantic rule-based complex event processing." In International Workshop on Rules and Rule Markup Languages for the Semantic Web, pp. 82-92. Springer, Berlin, Heidelberg, 2009.

S74 M. Ringsquandl, S. Lamparter, S. Brandt, T. Hubauer, and R. Lepratti. "Semantic-guided feature selection for industrial automation systems." In International Semantic Web Conference, pp. 225-240. Springer, Cham, 2015.

S75 J. U. Kietz, F. Serban, S. Fischer, and A. Bernstein. "Semantics Inside! But let's not tell the Data Miners: Intelligent Support for Data Mining." In European Semantic Web Conference, pp. 706-720. Springer, Cham, 2014.

S76 E. W. Patton, E. Brown, M. Poegel, H. De Los Santos, C. Fasano, K. P. Bennett, and D. L. McGuinness. "SemNExT: A Framework for Se-

mantically Integrating and Exploring Numeric Analyses." In SemStats@ ISWC. 2015.

S77 M. Spahn, J. Kleb, S. Grimm, and S. Scheidl. "Supporting business intelligence by providing ontology-based end-user information self-service." In Proceedings of the First international Workshop on ontology-Supported Business intelligence, p. 10. ACM, 2008.

S78 C. M. Keet,A. Lawrynowicz, C. dAmato, A. Kalousis, P. Nguyen, R. Palma, R. Stevens, and M. Hilario. "The data mining OPtimization ontology." Web Semantics: Science, Services and Agents on the World Wide Web, vol. 32, pp. 43-53, 2015.

S79 P. Panov, L. N. Soldatova, and S. Dzeroski. "Towards an ontology of data mining investigations." In International Conference on Discovery Science, pp. 257-271. Springer, Berlin, Heidelberg, 2009.

S80 E. Kharlamov, Y. Kotidis, T. Mailis, C. Neuenstadt, C. Nikolaou, Ö. Özcep, C. Svingos et al. "Towards analytics aware ontology based access to static and streaming data." In International Semantic Web Conference, pp. 344-362. Springer, Cham, 2016.

S81 D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, M. Ruzzi, and D. F. Savo. "The MASTRO system for ontology-based data access." Semantic Web, 2(1), pp. 43-53, 2011.

S82 D. Anicic,, S. Rudolph, P. Fodor, and N. Stojanovic. "Stream reasoning and complex event processing in ETALIS." Semantic Web, 3(4), pp. 397-407, 2012.

References

[1] G. Research, Gartner says advanced analytics is a top business priority (2015).

[2] H. Chen, R.H. Chiang and V.C. Storey, Business intelligence and analytics: from big data to big impact, *MIS quarterly* (2012), 1165–1188.

[3] O. Marjanovic, Improvement of Knowledge-Intensive Business Processes Through Analytics and Knowledge Sharing (2016).

[4] R. Espinosa, D. García-Saiz, M. Zorrilla, J.J. Zubcoff and J.-N. Mazón, Enabling non-expert users to apply data mining for bridging the big data divide, in: *International Symposium on Data-Driven Process Discovery and Analysis*, Springer, 2013, pp. 65–86.

[5] S.T. March and A.R. Hevner, Integrated decision support systems: A data warehousing perspective, *Decision Support Systems* **43**(3) (2007), 1031–1043.

[6] M. Magdon-Ismail, No free lunch for noise prediction, *Neural computation* **12**(3) (2000), 547–564.

[7] A. Abelló, O. Romero, T.B. Pedersen, R. Berlanga, V. Nebot, M.J. Aramburu and A. Simitsis, Using semantic web technologies for exploratory OLAP: a survey, *IEEE transactions on knowledge and data engineering* **27**(2) (2015), 571–588.

[8] P.A. Bernstein, D. Wecker, A. Krishnamurthy, D. Manocha, J. Gardner, N. Kolker, C. Reschke, J. Stombaugh, P. Vagata, E. Stewart et al., Technology and data-intensive science in the beginning of the 21st century, *Omics: a journal of integrative biology* **15**(4) (2011), 203–207.

[9] D. Fisher, R. DeLine, M. Czerwinski and S. Drucker, Interactions with big data analytics, *interactions* **19**(3) (2012), 50–59.

[10] P. Russom, Big data analytics, *TDWI best practices report, fourth quarter* **19** (2011), 40.

[11] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide (2000).

[12] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in knowledge discovery and data mining*, Vol. 21, AAAI press Menlo Park, 1996.

[13] F. Rabhi, M. Bandara, A. Namvar and O. Demirors, Big Data Analytics Has Little to Do with Analytics, in: *Service Research and Innovation*, Springer, 2017, pp. 3–17.

[14] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han et al., Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States, *Computing Research Association, Washington* (2012).

[15] Z. Milosevic, W. Chen, A. Berry and F. Rabhi, Real-Time Analytics, *Big Data: Principles and Paradigms* (2016), 39–61.

[16] F. Baader, *The description logic handbook: Theory, implementation and applications*, Cambridge university press, 2003.

[17] J.Z. Pan, S. Staab, U. Aßmann, J. Ebert and Y. Zhao, *Ontology-driven software development*, Springer Science & Business Media, 2012.

[18] K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson, Systematic Mapping Studies in Software Engineering., in: *EASE*, Vol. 8, 2008, pp. 68–77.

[19] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, ACM, 2014, p. 38.

[20] L. Yao and F.A. Rabhi, Building architectures for data-intensive science using the ADAGE framework, *Concurrency and Computation: Practice and Experience* **27**(5) (2015), 1188–1206.

[21] J. Taylor, Framing Requirements for Predictive Analytic Projects with Decision Modeling (2015).

[22] S. Shumilov, Y. Leng, M. El-Gayyar and A.B. Cremers, Distributed scientific workflow management for data-intensive applications, IEEE, 2008, pp. 65–73.

[23] G. Wang and Y. Wang, 3DM: domain-oriented data-driven data mining, *Fundamenta Informaticae* **90**(4) (2009), 395–426.

[24] S. Luján-Mora, J. Trujillo and I.-Y. Song, A UML profile for multidimensional modeling in data warehouses, *Data & Knowledge Engineering* **59**(3) (2006), 725–769.

[25] S. Luján-Mora and J. Trujillo, Physical modeling of data warehouses using UML, in: *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, ACM, 2004, pp. 48–57.

[26] H. Macià, V. Valero, G. Díaz, J. Boubeta-Puig and G. Ortiz, Complex Event Processing Modeling by Prioritized Colored Petri Nets, *IEEE Access* **4** (2016), 7425–7439.

[27] T. Berners-Lee, J. Hendler, O. Lassila et al., The semantic web, *Scientific american* **284**(5) (2001), 28–37.

[28] P. Ristoski and H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, *Web semantics: science, services and agents on the World Wide Web* **36** (2016), 1–22.

[29] J. Cardoso, M. Hepp and M.D. Lytras, *The semantic web: real-world applications from industry*, Vol. 6, Springer Science & Business Media, 2007.

[30] M.D. Lytras and R. García, Semantic Web applications: a framework for industry and business exploitation–What is needed for the adoption of the Semantic Web from the market and industry, *International Journal of Knowledge and Learning* **4**(1) (2008), 93–108.

[31] V.R. Benjamins, J. Davies, R. Baeza-Yates, P. Mika, H. Zaragoza, M. Greaves, J.M. Gomez-Perez, J. Contreras, J. Domingue and D. Fensel, Near-term prospects for semantic technologies, *IEEE Intelligent Systems* **23**(1) (2008).

[32] B.A. Kitchenham and S. Charters, Procedures for Performing Systematic Literature Reviews in Software Engineering, *EBSE Tech. Report version 2.3, EBSE-2007-01, Software Engin. Group, Keele Univ., Univ. of Durham, UK* (2007).

[33] E. Mendes, A systematic review of Web engineering research, IEEE, 2005, p. 10.

[34] N. Salleh, E. Mendes and J. Grundy, Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review, *IEEE Transactions on Software Engineering* **37**(4) (2011), 509–525.

[35] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: A practical guide*, John Wiley & Sons, 2008.

[36] K. Khan, R. Kunz, J. Kleijnen and G. Antes, *Systematic reviews to support evidence-based medicine*, Crc Press, 2011.

[37] L. Spencer, J. Ritchie, J. Lewis and L. Dillon, Quality in qualitative evaluation: a framework for assessing research evidence (2003).

[38] B. Elsevier, Scopus overview: What is it, 2008.

[39] H.O. Nigro, *Data Mining with Ontologies: Implementations, Findings, and Frameworks: Implementations, Findings, and Frameworks*, IGI Global, 2007.

[40] A. Guazzelli, M. Zeller, W. Chen and W. G., PMML: An Open Standard for Sharing Models, *The R Journal* **1**(1) (2009).

[41] T. Kliegr, V. Svátek, M. Šimunek, D. Stastnỳ and A. Hazucha, An XML schema and a topic map ontology for formalization of background knowledge in data mining, in: *IRMLeS-2010, 2nd ESWC Workshop on Inductive Reasoning and Machine Learning for the Semantic Web, Heraklion, Crete, Greece*, 2010.

[42] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog et al., The SSN ontology of the W3C semantic sensor network incubator group, *Web semantics: science, services and agents on the World Wide Web* **17** (2012), 25–32.

[43] G.O. Consortium et al., The Gene Ontology (GO) database and informatics resource, *Nucleic acids research* **32**(suppl 1) (2004), 258–261.

[44] J. Rogers and A. Rector, The GALEN ontology, *Medical Informatics Europe (MIE 96)* (1995), 174–178.

[45] H.M. Kim, M.S. Fox and M. Grüninger, An ontology for quality managementâĂŤenabling quality problem identification and tracing, *BT Technology Journal* **17**(4) (1999), 131–140.

[46] GeoVocab.org, GeoVocab, 2012, [Online; accessed 16-Feb-2017].

[47] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne et al., OWL-S: Semantic markup for web services, *W3C member submission* **22** (2004), 2007–04.

[48] R. Akkiraju, J. Farrell, J.A. Miller, M. Nagarajan, A.P. Sheth and K. Verma, Web service semantics-WSDL-S (2005).

[49] J. Kopeckỳ, T. Vitvar, C. Bournez and J. Farrell, Sawsdl: Semantic annotations for wsdl and xml schema, *IEEE Internet Computing* **11**(6) (2007).

[50] J. Domingue, D. Roman and M. Stollberg, Web service modeling ontology (WSMO)-An ontology for semantic web services, Jun, 2005.

[51] Markus Lanthaler, Hydra Core Vocabulary, 2017, [Online; accessed 16-Feb-2017].

[52] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik and J. Zhao, Prov-o: The prov ontology, *W3C recommendation* **30** (2013).

[53] A.I. Canhoto, Ontology-Based Interpretation and Validation of Mined Knowledge: Normative and Cognitive Factors in, *Data Mining with Ontologies: Implementations, Findings, and Frameworks: Implementations, Findings, and Frameworks* (2007), 84.

[54] M. Bandara, A. Behnaz, F.A. Rabhi and O. Demirors, From requirements to data analytics process: An ontology-based approach (In Press), in: *Proceedings of the 5th International Workshop on the Interrelations between Requirements Engineering and Business Process Management at BPM 2018*, Springer, 2018.

[55] A. Rajbhoj, V. Kulkarni and N. Bellarykar, Early experience with model-driven development of mapreduce based big data application, in: *Software Engineering Conference (APSEC), 2014 21st Asia-Pacific*, Vol. 1, IEEE, 2014, pp. 94–97.

[56] S. Ceri, E. Della Valle, D. Pedreschi and R. Trasarti, Mega-modeling for big data analytics, *Conceptual Modeling* (2012), 1–15.

[57] A.-L. Lamprecht, S. Naujokat, T. Margaria and B. Steffen, Semantics-based composition of EMBOSS services, *Journal of Biomedical Semantics* **2**(1) (2011), 5.

[58] N. Mehandjiev, F. Lécué, M. Carpenter and F.A. Rabhi, Cooperative service composition, in: *International Conference on Advanced Information Systems Engineering*, Springer, 2012, pp. 111–126.

[59] M. Bandara, F.A. Rabhi and R. Meymandpour, Semantic Model Based Approach for Knowledge Intensive Processes (In Press), in: *In Proceedings of International Conference on Software Process Improvement and Capabity Determination (SPICE)*, Springer, 2018.

[60] P. Panov, L.N. Soldatova and S. Džeroski, Generic ontology of datatypes, *Information Sciences* **329** (2016), 900–920.

[61] F. Rabhi, M. Bandara, A. Namvar and O. Demirors, Big Data Analytics Has Little to Do with Analytics, in: *Service Research and Innovation*, Springer, 2018, pp. 3–17.

[62] D. Withers, E. Kawas, L. McCarthy, B. Vandervalk and M. Wilkinson, Semantically-guided workflow construction in Taverna: the SADI and BioMoby plug-ins, in: *International*

Symposium On Leveraging Applications of Formal Methods, Verification and Validation, Springer, 2010, pp. 301–312.

[63] U. Dayal, M. Castellanos, A. Simitsis and K. Wilkinson, Data integration flows for business intelligence, in: *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, Acm, 2009, pp. 1–11.

[64] X. Zhang, Supporting on-the-fly data integration for bioinformatics, PhD thesis, The Ohio State University, 2007.

[65] D.E. Avison and G. Fitzgerald, Where now for development methodologies?, *Communications of the ACM* **46**(1) (2003), 78–82.