A Semantics-enabled approach for Data Lake Exploration Services

(Doctoral Symposium Paper)

Massimiliano Garda
(Supervised by: Valeria De Antonellis)
Dept. of Information Engineering, University of Brescia
Via Branze, 38 - 25123 Brescia (Italy)
m.garda001@unibs.it, valeria.deantonellis@unibs.it

***Abstract*—Ignited by the advent of Data Science, organisations are spending more and more resources in understanding their Big Data, attracted by the opportunity of turning them into actionable insights. Data Lakes have been proposed as repositories in charge of storing vast amount of heterogeneous data, regardless its structure, enabling the possibility of postponing transformation and analytical processes. In this context, Semantic Web technologies may be used to enable interoperability and improve data access, by providing Data Exploration Services. Starting from these premises, the goal of this paper is to describe a semantic approach apt to the compelling challenge of Data Exploration Services, aimed at personalising the exploration experience. The approach has been preliminary validated within a Smart City context, where aggregation of urban data, according to multiple perspectives through the definition of proper indicators, enables urban data exploration at different granularity levels for distinct categories of users.**

***Keywords*-data exploration services; data lake; semantic web; Big Data;**

I. Introduction

With the advent of Data Science, almost all organisations are spending time and resources in understanding their Big Data, attracted by the opportunity of turning them into actionable insights. For instance, stakeholders of a Smart City derive insights from a large variety of private and open data, ranging from energy consumption to weather forecasts, crime statistics, traffic data, to improve services delivered to citizens, predict future trends and enhance decision making procedures for ensuring citizens' wealth and security.

Data Lakes have been proposed as repositories in charge of storing this vast amount of heterogeneous data, regardless its structure, enabling the possibility of postponing transformation and analytical tasks. In the majority of the approaches, the main focus is foremost on how to build the Data Lake physical infrastructure and on the related data ingestion techniques. Novel approaches, focusing on the definition of services devoted to provide an effective and efficient exploration of data, are required. In this context, Semantic Web technologies may be used to enable interoperability and improve data access by providing Data Exploration Services. Starting from these premises, the goal of this paper is to describe a semantic approach apt to the compelling challenge of Data Exploration Services, aimed at personalising the exploration experience. The approach has been preliminary validated within a Smart City context, where aggregation of urban data, according to multiple perspectives through the definition of proper indicators, enables urban data exploration at different granularity levels for distinct categories of users. The approach has been presented in [4], where a more detailed discussion has been provided. The paper is organised as follows: in Section II an excerpt of the state of the art on Data Lake approaches is analysed; Section III presents the personalised exploration approach; Section IV discusses preliminary experiments; finally, Section V closes the paper, sketching future research directions.

II. Related Work

In its simplest definition, a Data Lake can be conceived as a storage repository that holds a vast amount of data in its native format, until it is needed. Within Data Lakes, heterogeneous data is stored and accessed; however, dealing with this data is cumbersome, time-consuming and inefficient, due to the different data models and file formats. Moreover, the access to this data is hindered by the variety of interfaces, services and applications. To address these issues, the idea is to leverage semantic information (also referred to as semantic *metadata*), which can mediate the Data Exploration process, modelling the knowledge needed to represent users' exploration demands.

A. Semantic representation

Given the versatility of Data Lakes in handling heterogeneous data, emerging works in this field have suggested to focus on the link between Data Lakes and semantic metadata, leading to the so-called *Semantic* Data Lakes. A broadly-recognised approach of semantic enrichment consists in linking data with external knowledge bases, thus performing data annotation [3]; other strategies, instead, focus on probabilistic techniques (e.g., labelling data through Fuzzy variables [9]). Moreover, to provide a context to data within the Data Lake, ontologies and knowledge graphs have been fostered in literature as a promising solution to offer a comprehensive view over the underlying data sources,

978-1-7281-3851-0/19/$31.00 ©2019 IEEE
DOI 10.1109/SERVICES.2019.00091

modelling their interrelationships and dependencies [2], [8]. Likewise, in [14] knowledge graphs store data fragments, represented as different typologies of nodes inside the graph. In the former approaches, there is no mention about fostering such knowledge to achieve Data Exploration, from the users' viewpoint. Sharing some issues with the proposal of this paper, the approach presented in [5] harnesses thematic views upon the datasources of the data lake, but again personalisation aspects are treated only at a high level of abstraction, without a fully-fledged semantic support. Lastly, a semantic data platform has been presented in [11], adopting a flexible data ingestion pipeline, but exploration aspects have been left in the background.

B. Data Exploration support

Several attempts have been made to provide Service Oriented Architectures apt to support users in Data Exploration tasks when dealing with Data Lakes. In [3] proper services are implemented on the top of a storage layer, with the aim of assembling a knowledge graph compliant with the exploration request issued by the user. Similarly, in [1] different services are envisaged in order to assist users in combining and integrating data stored within a data lake. Focusing on the architectural perspective, authors in [6] design a modular and scalable system, composed of a set of services, yielding to a framework wherein users may invoke analytical modules on-demand. Beyond implemented services, in [12] personalisation aspects, concerning the healthcare domain, are taken into account to suggest recommendation targeted to clusters of users.

In all these approaches, the focus is more on the visualisation and analytical aspects, instead of proposing techniques to accomplish user's exploration demands, thus suggesting proper exploration directions.

III. Personalised approach for Data Lake exploration

In this section, the personalised approach for Data Lake exploration is introduced specifying its main pillars, namely: (a) Data Lake representation; (b) the ontology apt to provide the knowledge required to model exploration scenarios, exploration dimensions and indicators and (c) the services (together with the related modules) interacting with the ontology and responsible for providing Data Exploration support to users.

A. Data Lake Representation

Herein, a Data Lake is represented a set of datasources, namely $DL = \langle \mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n \rangle$. A datasource $\mathcal{S}_i$ in DL can be formalised by the tuple $\mathcal{S}_i = \langle \mathcal{A}_i, \mathcal{O}_i, \mathcal{M}_i \rangle$ where: (i) $\mathcal{A}_i$ is a set of attributes; (ii) $\mathcal{O}_i$ is a set holding attribute-value pairs, with attributes belonging to $\mathcal{A}_i$, representing the content of the datasource and (iii) $\mathcal{M}_i$ is a set containing attribute-value pairs, with attributes belonging to $\mathcal{A}_i$, representing metadata associated to the datasource.

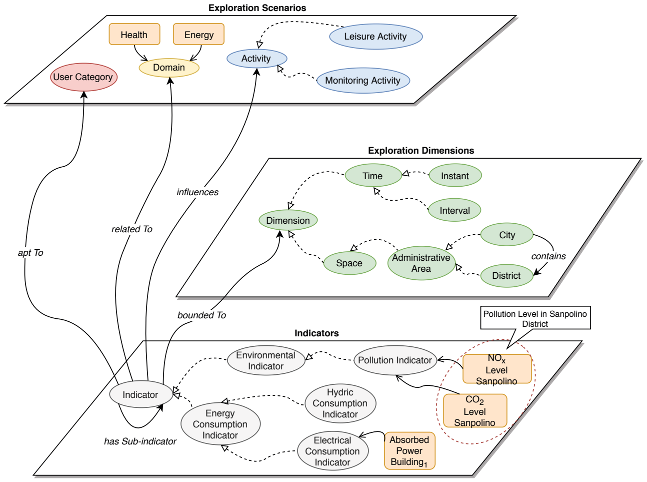

Figure 1. Representation over three layers of ontology-based exploration (exemplification for the Smart City domain).

For what concerns the Smart City domain recalled in Section I, an example of datasource $\mathcal{S}_i$ may be a stream of measures generated by a sensor, measuring particulate concentration in a specific district. In this context, physical quantities measured by sensors are the values of proper indicators, which are designed on top of these datasources. Coherently, candidate attributes for the set $\mathcal{A}_i$ would be timestamps (useful to temporally locate data), sensor configuration (e.g., names of measured quantities, geographical area the sensor belongs to) and so forth. Sensor configuration attributes are amenable to be considered as metadata, and therefore appropriate to be included in the set $\mathcal{M}_i$. Hence, the content of the set $\mathcal{M}_i$ can be organised by distinguishing diverse typologies of metadata, apt to serve different purposes, complying with the general classification given in [10]. Specifically, it can be: (i) *technical*, retaining information related to data format and schema; (ii) *operational*, including information automatically generated during data processing (e.g., provenance attribution) and (iii) *business*, concerning business rules involving upper/lower limits of a particular field, data domain and integrity constraints and so forth.

Hereafter, a particular concern will be given to business metadata, as it may be leveraged to model user-tailored exploration scenarios.

B. Ontology-based Exploration

As previously remarked, Semantic Web technologies may be employed to enable interoperability and improve data access. In particular, ontologies are advocated whenever it is necessary to move the focus towards the meaning of data, modelling the domain knowledge through concepts and relationships. To provide a semantic description for concepts and relationships required to model users' personal exploration demands, *exploration scenarios* have been considered. Exploration scenarios can be conceived as stereotype data exploration patterns, with the goal of easing analytical tasks, explicitly envisaged for different user's categories. For instance, in the Smart City domain, an example of exploration scenario would be related to building monitoring (performed by building administrators) for electrical consumption. In order to model exploration scenarios and the knowledge related to indicators, an ontology has been constructed by following best practices in ontology designing, fully detailed in [4]. Such ontology is organised over three conceptualisation layers (Figure 1), whose description is given in the following (starting from the topmost).

1) Exploration Scenarios: Owing to the wide variety of urban data that can be explored, indicators exploration can be personalised taking into account user's category (e.g., citizen, building administrator, Public Administration), activities practised by users and domains of interest (e.g., environment, health). These elements are the ingredients of exploration scenarios and, at the same time, constitute user's profile data.

2) Exploration Dimensions: Indicators can be segmented according to several dimensions. Specifically, temporal dimensions are exploited to supply a temporal reference to the values provided by an indicator. Differently, spatial dimensions are suitable to define the spatial coverage for an indicator, which follows the inherent city spatial organisation. Over the latter set of dimensions, it is also possible to define containment relationships (e.g., to state that a city includes several districts).

3) Indicators: Indicators can be hierarchically organised (e.g., to specialise consumption indicators into hydric and electrical). In the ontology, an indicator is apt to one or more user's categories (e.g., citizens, energy providers, Public Administration), is related to one or more domains (e.g., health, environment, energy) and influences activities performed by users (e.g., leisure activities, building monitoring activities).

C. Semantics-enabled Exploration Services

Figure 2 depicts the architecture whereby semantics-enabled Data Exploration services have been deployed.

Hereafter, the *Semantics Service*, lying in the Semantic Layer of the architecture, is presented through the description of its constituent modules, without lingering much on the details of other services (namely, *Request Processor Service* and *Visualisation Services*), as their technical description is out of the scope of this paper.

Users can issue queries as a set of keywords, that will undergo a Wordnet-based disambiguation process and then matched with ontology objects, according to renowned techniques [13], in order to provide support to the request formulation without demanding a detailed knowledge of ontology structure by the users. Given a user's query, the semantics-enabled exploration is articulated over two steps: (i) indicators are properly recommended taking into account both explicit requirements of the user, as expressed in a search request, and the users profile; (ii) recommended indicators are exploited as a starting point to set up interactive exploration of urban data.

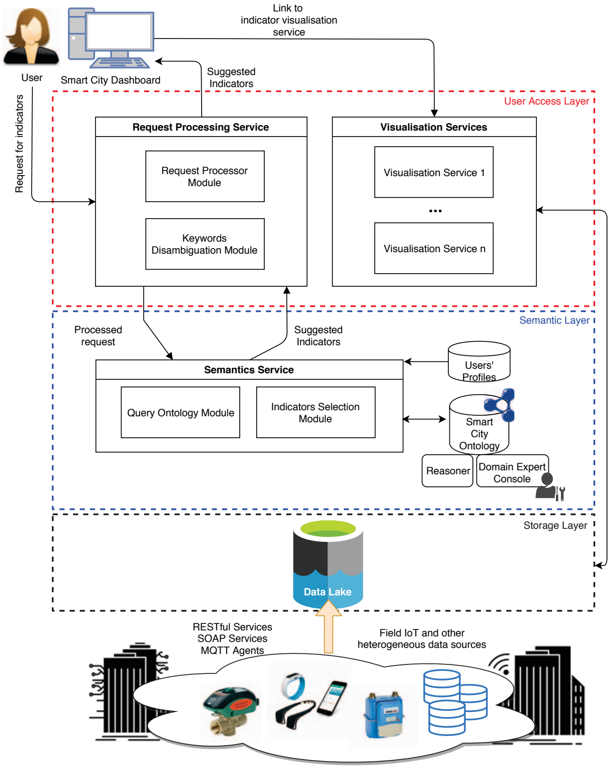

Figure 2. Overview of the semantics-enabled Data Lake exploration Service Oriented Architecture.

Query Ontology Module: This module operates on the ontology, presented in Section III-B, to retrieve and select indicators compliant with the user's request. Specifically, the module is in charge of issuing *selection* and *boolean* queries over the ontology (in SPARQL language). The former ones are used to retrieve individuals from the ontology (e.g., to collect the set of indicators apt to citizens) whereas the latter ones are useful to query the ontology to get true/false responses (e.g., to check whether an indicator is available for citizens). These queries can benefit from the underlying reasoning engine, to enrich their results when computing hierarchies (e.g., to build the lineage of an indicator, up to its topmost ancestor) or to infer additional background knowledge.

Indicators Selection Module: This module filters out retrieved indicators which are not compliant with user's profile. The selection procedure (thoroughly detailed in [4]) considers user's category, activities and places where he/she acts within the Smart City. Depending on the profile, citizens can select indicators concerning their apartment only, building managers can select indicators on their administered buildings only, energy managers can select indicators that only concern the workplaces they are responsible for, etc. Each suggested indicator has a link which triggers a proper visualisation service inside the platform, that enables to display indicator values onto the Smart City Dashboard, accessing the data within the Data Lake.

IV. Experimental Validation

Preliminary experiments on the proposed approach aimed at demonstrating its effectiveness in supporting indicators selection and have been discussed in [4]. In this respect, two kinds of requests have been considered: (A) requests where the user specified a set of keywords in order to identify desired domains and indicators, and the user's profile does not contain any activity or preferential indicator; (B) requests where the user presents a richer profile (containing category, activities and preferential indicators), but specifies keywords that only correspond to domain individuals. The illustrated ontology-based approach has been compared against a keyword-based search, where semantic disambiguation techniques have been applied to the set of keywords [13], but semantic relationships within the ontology have not been exploited. Average precision and recall values for the keyword-based search are equal to 0.49 and 0.97 for type A (0.33 and 0.27 for type B, resp.), whereas for the ontology-based search they are equal to 0.99 and 0.98 for type A (0.94 and 0.93 for type B, resp.). Candidate indicators selection average execution time for type A is about 2559 ms, whereas for type B is about 1325 ms. Since both the compared approaches use keywords disambiguation techniques and the same keywords have been used during tests, difference in average precision and recall is due to the knowledge structure in the ontology.

V. Conclusion and Future Work

Given the dissertation of the previous sections, ongoing steps are directed to refine the approach described in this paper, originated from [4]. Specifically, the contributions will be twofold. On the one hand, getting inspired by the principles of [7], the legacy Storage Layer of the architecture will be organised into *zones*, each of them fulfilling a specific role in the broader Data Exploration process. On top of this Service Oriented Architecture, proper modules will be deployed to perform *data summarisation* (particularly advised for streaming datasources) and *data relevance evaluation* (to focus the exploration only on a subset of data deemed significant/worthy to be analysed by users). On the other hand, the ontology formerly described will be at the basis of a *knowledge graph* extraction process, in which semantic information will be linked to the (relevant) data in the Data Lake, for exploration purposes. Overall, further experiments will be carried on, enhancing the ontology by considering additional relationships between indicators (e.g., to assert that two or more environmental indicators must be jointly monitored due to their harmful impact on the ecosystem) as a basis to dispense fruitful advice for users.

References

[1] H. Alili, K. Belhajjame, R. Drira, D. Grigori, and H. H. B. Ghezala. Quality Based Data Integration for Enriching User Data Sources in Service Lakes. In *2018 IEEE International Conference on Web Services (ICWS)*, pages 163–170, 2018.

[2] A. Alserafi, A. Abelló, O. Romero, and T. Calders. Towards information profiling: Data lake content metadata management. In *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, pages 178–185, 2016.

[3] A. Beheshti, B. Benatallah, R. Nouri, and A. Tabebordbar. CoreKG: a Knowledge Lake Service. *PVLDB*, 11(12):1942–1945, 2018.

[4] D. Bianchini, V. De Antonellis, M. Garda, and M. Melchiori. Exploiting Smart City Ontology and Citizens' Profiles for Urban Data Exploration. In *26th Int. Conference on Cooperative Information Systems (CoopIS)*, pages 372–389, 2018.

[5] P. Lo Giudice, L. Musarella, G. Sofo, and D. Ursino. An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences*, 478:606–626, 2019.

[6] J. Herman, H. Herman, M. J. Mathews, and J. C. Vosloo. Using big data for insights into sustainable energy consumption in industrial and mining sectors. *Journal of Cleaner Production*, 197:1352–1364, 2018.

[7] B. Inmon. *Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump*. Technics publications, 2016.

[8] A. Maccioni and R. Torlone. Crossing the finish line faster when paddling the Data Lake with KAYAK. *PVLDB*, 10(12):1853–1856, 2017.

[9] B. Malysiak-Mrozek, M. Stabla, and D. Mrozek. Soft and Declarative Fishing of Information in Big Data Lake. *IEEE Transactions on Fuzzy Systems*, page 1, 2018.

[10] A Oram. *Managing the Data Lake*. O'Reilly, Sebastopol, 2015.

[11] A. Pomp, A. Paulus, A. Kirmse, V. Kraus, and T. Meisen. Applying semantics to reduce the time to analytics within complex heterogeneous infrastructures. *Technologies*, 6(3):86, 2018.

[12] S. Rangarajan, H. Liu, H. Wang, and C.-L. Wang. Scalable Architecture for Personalized Healthcare Service Recommendation Using Big Data Lake. In *Service Research and Innovation*, pages 65–79. 2015.

[13] J. A. Royo, E. Mena, J. Bernad, and A. Illarramendi. Searching the web: From keywords to semantic queries. In *Third Int. Conf. on Information Technology and Applications (ICITA 2005)*, pages 244–249, 2005.

[14] C. Walker and H. Alrehamy. Personal Data Lake with Data Gravity Pull. In *Proc. of IEEE 5th Int. Conf. Big Data and Cloud Computing*, pages 160–167, 2015.