A Modelling Framework for Evidence-based Public Health Policy Making

Marios Prasinos, Ioannis Basdekis, Marco Anisetti, Senior Member, IEEE, George Spanoudakis, *Member, IEEE,* Dimitrios Koutsouris, Senior Member, IEEE and Ernesto Damiani, Senior Member, IEEE

***Abstract*— It is widely recognised that the process of public health policy making (i.e., the analysis, action plan design, execution, monitoring and evaluation of public health policies) should be evidenced based, and supported by data analytics and decision-making tools tailored to it. This is because the management of health conditions and their consequences at a public health policy making level can benefit from such type of analysis of heterogeneous data, including health care devices usage, physiological, cognitive, clinical and medication, personal, behavioural, lifestyle data, occupational and environmental data. In this paper we present a novel approach to public health policy making in a form of an ontology, and an integrated platform for realising this approach. Our solution is model-driven and makes use of big data analytics technology. More specifically, it is based on public health policy decision making (PHPDM) models that steer the public health policy decision making process by defining the data that need to be collected, the ways in which they should be analysed in order to produce the evidence useful for public health policymaking, how this evidence may support or contradict various policy interventions (actions), and the stakeholders involved in the decision-making process. The resulted web-based platform has been implemented using Hadoop, Spark and HBASE, developed in the context of a research programme on public health policy making for the management of hearing loss called EVOTION, funded by the Horizon 2020.**

***Index Terms*— model driven data analytics, evidence-based health policy making, ontologies, public health policy;**

I. Introduction

The effective management of various health conditions depends on and requires appropriate public health policies (PHP) [1]. Health policy affects the access to health care services (e.g., health check-ups, health care device adjustments, provision of related rehabilitation services), medication and supportive devices. It also affects the provision of services related to screening for prevention of disease, early diagnosis and treatment, long-term management of chronic diseases and disabilities, and as such its effectiveness should been monitored by means of bringing all data together to analyse and interpret the progress and the final outcome of this implementation [2]. In this context, the World Health Organization (WHO) defines health policy as the "decisions, plans, and actions that are undertaken to achieve specific health care goals within a society"[3], the making of which consists of four key stages: i) Situation analysis; ii) Development of action plan; iii) Implementation and monitoring of programme; and iv) Programme evaluation.

It is widely recognised that the life cycle of PHP making and its aforementioned stages (in brief, analysis, action plan design, execution, monitoring and evaluation of a PHP) should be evidenced based (onwards EBPHP) [4]. This is because the management of health conditions and their consequences at a public health policy making level can benefit from the analysis of heterogeneous data, including health care device usage, physiological, cognitive, clinical and medication, personal, behavioural, lifestyle data, occupational environmental and several other available types of data. According to F. Rajabi, ultimate goal of a health system is community health promotion in an equitable manner, and as such evidence is required so that policy makers be able to assess more objectively the effectiveness of a policy in question [5]. The analysis of data can enable the investigation of cause-symptom effects, comorbidities and contextual factors, including social, behavioural and economic and lifestyle factors, that may relate and/or affect disease management. Yet in terms of applicability, forming models, associating goals to calculated indexes and assessing analytics results for such data heterogeneity (i.e., policy making process) is a rather mentally demanding and time-consuming task, while at the same time the expectation is for faster, well defined, and user-friendly decision-making processes. In addition, the potential impact and the greater benefit of any intervention in question is weighed against economic objective factors as well such as the overall cost, thus concrete quantitative evidence derived from this wealth of information should be backed by rigorously synthesised research analytics. As an instance of this, it is predominately assumed that the right policy prescription can be derived from relevant research evidence [6]. Thus, the outcomes of such

Manuscript received March 18, 2021. This work has been partly supported by the EU-funded project EVOTION (grant no H2020-727521).

Marios Prasinos, City, University of London, EC1V 0HB, London, United Kingdom (phone: 0030 6932907702; e-mail: Marios.Prasinos.1@city.ac.uk).

Ioannis Basdekis, City, University of London, EC1V 0HB, London, United Kingdom (e-mail: Ioannis.Basdekis@city.ac.uk).

Marco Anisetti, Dipartimento di Informatica Università degli Studi di Milano, 26013, Milano, Italy (e-mail: Marco.Anisetti@unimi.it).

George Spanoudakis, City, University of London, EC1V 0HB, London, United Kingdom (e-mail: G.E.Spanoudakis@city.ac.uk).

Dimitrios Koutsouris, Biomedical Engineering Laboratory, National Technical University of Athens, 15773, Athens, Greece (e-mail: dkoutsou@biomed.ntua.gr).

Ernesto Damiani, Dipartimento di Informatica Università degli Studi di Milano, Milano, Italy (e-mail: Ernesto.Damiani@unimi.it).

analysis can enable the stratification of related risks and effects to the patients, and – through correlation with other economic, social and physical constraints – enable the development of comprehensive/holistic interventions to the management of health conditions and the overall well-being of patients.

Worth's mentioning that end-user configurable data analytics can also help exploring missing, under or over-estimated values of specific medical interventions, thus cleaning the input data prior of analysing and deriving conclusions (pre-processing analytics). It is therefore of enormous importance that orchestrating tasks executions (which to be executed first, in parallel, waiting for a correlated Task to be finished, etc.) via methods for efficient scheduling of tasks jobs are considered requirements for an ideal framework for processing of such large datasets by using parallel and distributed programming approaches [7].

Despite such potential benefits, however, at present the PHP making is mainly supported by guidelines and policy-modelling tools which do not integrate and use data analytics as such, and decision-making tools based on the outcomes of such analytics.

In this paper, we present a novel approach to support the PHP making and an integrated platform for realising this approach. Our approach is model-driven, hence it is based on end-user administered public health policy decision making (onwards PHPDM) models. These models steer the public health policy decision making process by defining the data that need to be collected, the ways in which they should be analysed in order to produce the evidence useful for PHP making, how this evidence may support or contradict various policy interventions (actions), and the stakeholders involved in the decision-making process. The platform that instantiates the proposed model-driven specification that supports the management of policies is based on the use of big data analytic technologies [8], as in order to address the current challenges related to the existence of very large, structurally heterogeneous, and fast-growing healthcare data sets, which may be collected from different and dynamically evolving data sources, through the use of heterogeneous devices and technologies and are often relevant in PHP decision making processes ([9], [10]).

A preliminary version of the specification of the proposed modelling framework was presented in [11]. Here, an extended version of this specification is presented, along to execution results of a use case scenario, in the context of Hearing Loss (HL) management. The rest of this paper is structured as follows. Section II describes related work. Section III presents the basis of a PHPDM framework. Section IV presents the ontology-based scheme for specifying PHPDM models, Section V presents an example PHPDM model, while Section VI shows how it can be executed to realise our approach. Finally, Section VI presents experimental evaluation and discussion.

II. Related Work

The formation of a public health policy (PHP) making usually entails four main stages: (i) situational analysis; (ii) development of action plan; (iii) implementation and monitoring of programme; and (iv) programme evaluation in long/medium term [12]. Hence, it would be beneficial for PHP makers to be provided with an integrated working environment that encapsulates and assists the implementation of all those stages, by the provision of necessary tools on this path of analysis and assessments during the decision-making process. Currently, these stages are supported fragmentally by questionnaire-based assessments (e.g., [13], [14], [15], [16], [17], [18]), survey-based frameworks (e.g., a framework for evaluating adult hearing services using outcomes relevant to service users [10]), and training platforms [19]. Most of those frameworks introduce performance indicators, created together with service descriptors with input from stakeholders, giving the latter the power to determine which of those indicators and descriptors are most useful to them. Still, do not exploit advantages of real-time big data analytics nor incorporate modern analytics capabilities (from basic analytics to complex ML ones) and decision support (i.e., weighted decision criteria) in a user-friendly manner. do not exploit advantages of real-time big data analytics (e.g., [16], [17], [18]) nor incorporate data analysis capabilities (from basic analytics to complex ML ones) and decision support (i.e., weighted decision criteria). A representative example of a questionnaire-based assessment is the CFHI Assessment Tool™[18], to support EIHPM. This tool aims in guiding health organizations toward making the changes needed to become high-performing ones via measuring improvements in patient care, population health and value for money. CFHI although covers as well all stages(i)-(iv) mentioned, still implements simple guidelines to identify and incorporate evidence in health policy formulation. Although these frameworks assist policy makers on directing investigations of the literature as part of PHP decision, we argue that they suffer from many shortcomings. WHOs Ear and Hearing Care Situation Analysis Tool (EHCSAT) can be considered as tool to describe and assess the need for ear and hearing care services, but its static nature does not cater one of the most predominate cross-sectoral technological features, the analysis of big data [13]. Nowadays the policy making via the effective use of big data analytics has nurtured enthusiasm for evidence-based analysis and assessments. Specialized articles for forecasting trends in economic policies (e.g., [20]), defence policies (e.g., [21]), and many other policy sectors (e.g., in the field of security a review by [22]), reemphasize the potential utility derived from such analysis. Notably, although policy making process have always owned and processed large (in terms of volume) portions of data, still the plethora currently collected from different sources provides opportunities to discover and extract knowledge in places that have never been tested or previously identified as potential source of information. Big Data engines are largely used in healthcare related application. Specifically for healthcare scenario, [23] discuss benefits and outshine architectural framework and methodology. [24] proposes a more patient-centric healthcare system built on Cloud and Big Data Analytics, while [25] similarly focused on assisted leaving healthcare system, supported by Big Data processing. The work by B. Fabian et al focused at architectural level proposing Cloud and Big Data base system where the security of the health-related data is protected by a number of encryption mechanisms. [26]. Our previous work in [27] presents a simplified policy making approach with some usage examples without detailing the connection between the policy and the evidence obtained via big data analytics. This paper fills in this gap providing a powerful modelling

framework. Last but not least, one has to consider the execution complexity of matching input datasets, deploy and run according to a time series that has inherent limitations (e.g., re-executions according to predefined time or data related criteria, smart re-utilisations of outputs of one execution as input to another) without manual intervention. In this dimension, proposed solutions demonstrate limited applicability and extension possibilities for the community [28].

The field of HL (in particular interest of the European Union-funded project EVOTION [29], in the context of which the introduced modelling framework implemented) can benefit from the use of big data generated by HAs, associated with medical records and well-being data to provide evidence not only for improving hearing but also to inform decisions at the population level [30]. Consequently, we argue that at the level of HL a holistic management requires a well-defined specification, and underlying tools for investigating appropriate public health policies for HL any aspect associated with treatment: prevention; early diagnosis; long-term treatment and rehabilitation; detection and prevention of cognitive decline; protection from noise; socioeconomic inclusion of HL patients, and others. Tools to support complex analysis of heterogeneous big data (e.g., HA usage, noise (TTS) episodes, audiological, physiological, cognitive, clinical and medication, personal, occupational and environmental data), time and data and prioritised constrained executions, and in the end a user-friendly interactive environment to assist the knowledge extraction and the presentation of the evidence-based assessments.

In the case presented onwards, an extended specification and tools supporting the EBPHP unfolds. This architecture is much in line with the work of big data engine presented in [31], where the engine defined supports simple deontic logic policies.

III. Related Ontologies

In this section we present the existing ontologies in the domains of policy making, data mining and statistics we got inspired from to build our ontological modelling framework.

A. Policy Making

[31] has developed an ontology-based approach for modelling public policies and managing them across their entire life cycle. This approach has been developed with the intention to support policy modelling and management in a collaborative manner involving interactions between different stakeholders involved in such activities, and in particular cases where policy modelling involves collaboration between different government stakeholders (i.e., G2G collaboration). In addition, a domain independent (referred to by the author as "horizontal") ontology was proposed for modelling public policy processes, which – according to the author – could be used for governmental policy formation processes in different do-mains subject to extensions of the core horizontal ontology with domain specific ontologies. The modelling of policies in this approach is based on five core ontological concepts. These are: the *issue* (i.e., the problem to be solved or goal to be achieved by the policy); the *alternatives* (i.e., the alternative directions of action/ways in which the issue(s) can be addressed); the *positions* that different stakeholders may express on different alternatives (positions can be support or object to alternatives); the *preferences* that different stakeholders may express on different positions to indicate their relative importance; and the *criteria* that will be used to reach decisions.

Our ontological framework is based on the approach introduced in [11] for the policy stakeholders and decision-making process's part. We have also been inspired by G2G ontology introduced by Loukis et al[32], but created our own approach for the policy aims, objectives and actions module.

B. Data Mining

A reference modular ontology for the domain of data mining OntoDM proposed by Panov[33], was directly motivated by the need for formalization of the data mining domain. The OntoDM ontology is designed and implemented by following ontology known practices and design principles. Its distinguishing feature is that it uses Basic Formal Ontology [34] (BFO) as an upper-level ontology and a template, a set of formally defined relations from Relational Ontology [35] (RO) and other state-of-the-art ontologies, and reuses classes and relations from the Ontology of Biomedical Investigations [36] (OBI), the Information Artifact Ontology [37] (IAO), and the Software Ontology [35] (SWO). This ensures compatibility and connections with other ontologies and allow cross- domain reasoning capabilities.

The main ingredient in the process of data mining is the data. In OntoDM-core, they model the data with a data specification entity that describes the datatype of the underlying data. For this purpose, they import the mechanism for representing arbitrarily complex datatypes from OntoDT ontology [38].

We were inspired by OntoDM-core for the data mining part of our ontology. OntoDM-core is far more complex than our needs, but it was a good basis for modelling our approach.

C. Statistics

STATO [39] is a general-purpose statistics ontology, whose aim is to provide coverage for statistical processes such as statistical tests, the conditions of their application, and the information needed or resulting from statistical methods, such as probability distributions, variable, spread and variation metrics. The specific ontology also covers aspects of experimental design and description of plots and graphical representations commonly used to provide visual cues of data distribution or layout and to assist review of the results.

STATO provides textual definitions for all terms, as well as formal definitions for most of the terms allowing automatic classification, for example, categorising the statistical methods depending on the nature of the variables used as input, the conditions and their domain of application.

In our ontological framework we were based on STATO and used the classes that describe the statistical algorithms of our approach.

IV. Public Health Policy Decision Making

The development of the PHPDM model specification framework (introduced in [11]) has been driven by the need to provide a framework enabling the specification of:
(a) the overall goal and the specific objectives that public policy needs to address in a given area of health intervention;

(b) the range of possible actions (interventions) through which the goals and objectives of the policy can be achieved;
(c) the evidence that needs to be gathered and analysed in order to make informed and plausible decisions about the actions (interventions) that need to be undertaken (made) as part of the policy;
(d) the processes for analysing and establishing the validity of this evidence;
(e) the stakeholders who will consider the evidence and decide which actions (interventions) should be undertaken (made);
(f) the criteria that should be used to make decisions on the basis of the identified evidence.

The proposed PHPDM model specification framework has been defined as an ontology using the W3C Web Ontology Language (OWL) [40]. This has been due to the fact that OWL is an established and widely used ontology modelling framework that provides a framework for defining ontologies with formal model theoretic semantics. Hence, the use of OWL has enabled us to specify the PHPDM framework in a formal manner, enabling the automated and formal reasoning about PHPDM models, using several tools that are available for this purpose (e.g., Protégé [41]). Furthermore, OWL comes with several standardised syntaxes for defining ontologies, which are based or can be transformed to Resource Description Framework (RDF) [42], as for example the Manchester [43], Turtle [44], RDF/XML [45] and OWL2/XML [46] syntax. Hence, the definition of PHPDM models in OWL makes them easily ex-changeable across different tools and applications.

The PHPDM framework addresses the needs (a)-(f) identified above. To do so, it introduces modelling constructs defined as OWL classes, that enable the specification of PHPDM model elements covering (a)-(f). The very modelling of these constructs in OWL constitutes an element of formalisation. However, it is not the only element of formalisation. The definition of the PHPDM framework is also based on axioms, which are introduced to restrict the possible use of its constructs, where necessary. In the following section, we provide an overview of the top-level OWL classes and structure of the PHPDM framework, discuss the different types of users that we envisage for the framework, outline the semantic foundations of OWL that apply and give semantics to it, and introduce the syntax that we have used for the framework.

V. Specification of PHPDM Models

The definition of the PHPDM framework and the individual PHPDM models specified in it is ontology based. More specifically, the PHPDM framework is defined as a set of classes in the ontology modelling framework OWL and PHPDM models are specified as interrelated instances of these classes (i.e., objects which instantiate the classes of the framework and are related by instances of the relationships defined in the framework). This is because OWL provides a modelling framework with clear semantic foundations, providing a solid basis for processing (i.e., querying, drawing inferences, interpreting and executing) the models defined in it.

Conceptually, the PHPDM framework can be broken down into 3 modules:

- **The policy module** (referred to as "Module 1" in the rest of the manuscript): This module includes classes that specify the overall goal and objectives that a health policy that needs to be formed should address, and the actions (interventions) that will be needed to realise the policy.
- **The policy making module** (referred to as "Module 2" in the rest of the manuscript): This module includes the classes that specify the stakeholders who participate in the decision-making process and the positions that they may express.
- **The data analytics and evidence module** (referred to as "Module 3" in the rest of the manuscript): This module includes classes that specify the data that will need to be analysed to produce evidence aiding the making of policy decisions, the forms of analysis that should be applied to these data, and the criteria that should be used to assess whether the evidence generated from the data is sufficient in supporting actions.

Figure 1 shows the top-level classes and relationships of the ontology that constitutes the PHPDM framework. The figure shows these classes and their relationships as a UML [47] class diagram. It should be noted that the use of UML to present the ontology that defines the PHPDM framework has been adopted merely to enable the visual presentation of the framework and does not constitute part of the definition of the framework.

The class *PolicyModel* in Figure 1 is the class that can be used to specify PHPDM models. As shown in the figure, each *PolicyModel* has a general *Goal* (i.e., a possibly non-measurable target that it aims to address). The *Goal* of a policy is expressed at a generic level and its achievement requires addressing concrete *Objectives*. *Objectives* are measurable policy targets that can be addressed by *PolicyActions*.

A *PolicyAction* presents a possible way of addressing one of the objectives of the policy. A *PolicyAction* can act as pre-requisite, as dependency or as dependant to other policy actions. Policy actions may need to be applied as alternative (i.e., mutually exclusive) or complementary means for realising the objectives of a policy. The possible ways of applying actions are specified by the policy model. More specifically, in cases where actions need to be applied as alternatives, the model must describe them as such.

A. Example 1 - PolicyModel

Policy actions reflect the key decisions that may be made in PHPDM process. These decisions need to be explored on the basis of evidence arising from the analysis of data. To express this, in the PHPDM framework ontology each policy action is associated with a *Criterion* that determines the circumstances under which the evidence arising from data analytics would support the action. A criterion is specified by a *LogicalExpression* over the outcomes of a *DataAnalyticsWorkflow*.

With regards to the type of processing that they perform upon their input data set(s), data analytics tasks can be distinguished into *StatisticalAnalysisTasks* (i.e., tasks that carry out some statistical analysis upon the data), *DataMiningTasks* (i.e., tasks that carry out some data mining analysis upon the data), *SocialMediaAnalyticsTasks* (i.e. tasks that carry out some analysis of social media data), *SimulationTasks*, (i.e. tasks that carry out analysis of simulating-synthetic data), *TextMiningTasks* (i.e.

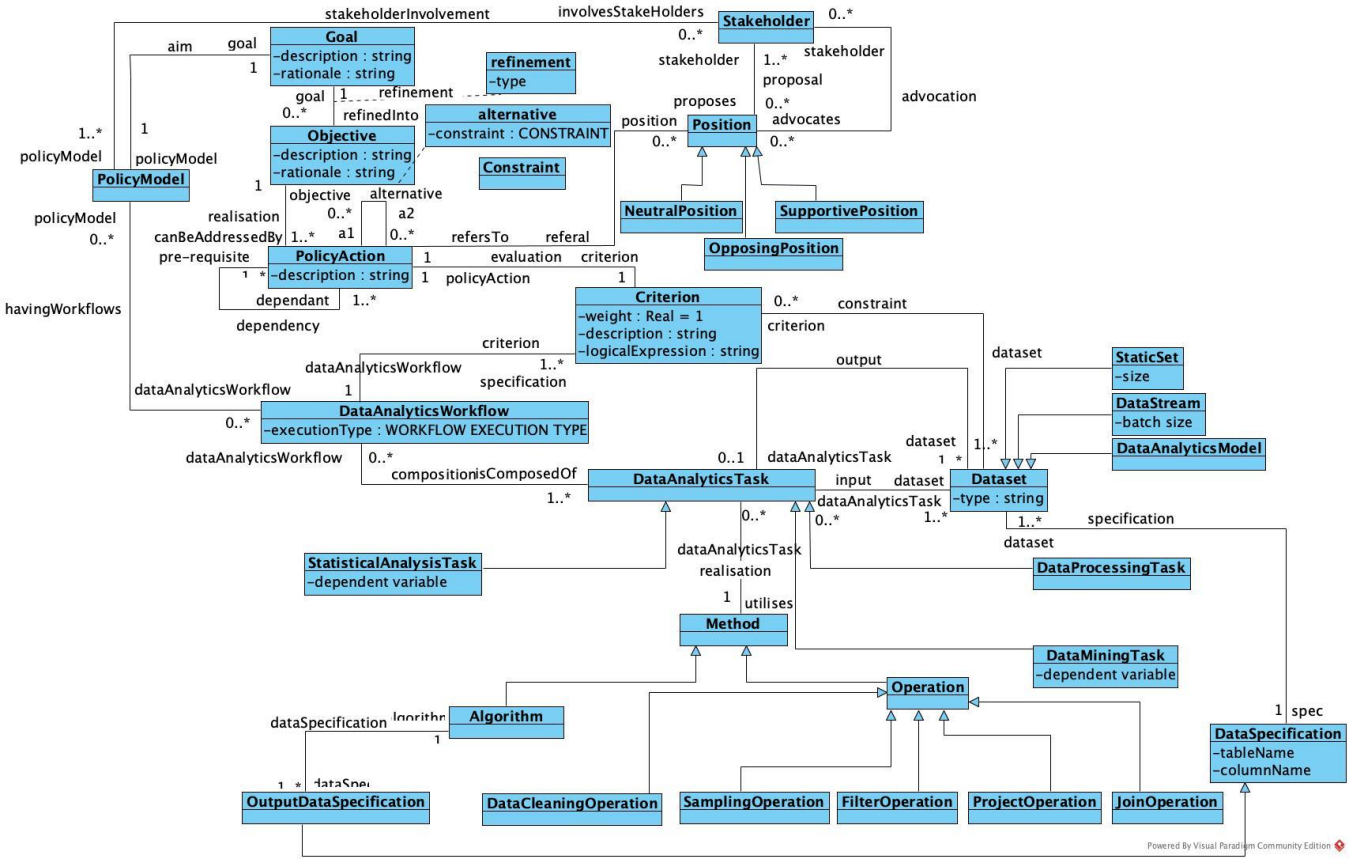

Fig. 1. Main Classes of the Ontology Framework.

tasks that perform text mining techniques for analysis of the literature) and *DataProcessingTasks* (i.e., tasks that perform some pre-processing over the data that is required prior to the analysis such as data cleaning or data joining over correlating factors). Each data analytics task utilizes a *Method*, which can be an *Algorithm* or an *Operation*. Each algorithm comes with an *OutputDataSpecification* (i.e., the form of the output data). All types of tasks utilize algorithms, except data processing tasks which utilize operations. Thus, *DataProcessingTasks* such as *Filtering* are used for pre-processing (e.g., spot and clean missing data, find, fill/remove outliers), prior of the execution of statistical and AI analysis supported (i.e., Basic stats (Mean, Variance, Min, Max, NormL1, NormL2, Median), Distribution, Linear Regression Model, ANOVA, K-means clustering, Linear Regression, F-test, Decision tree classification, Principal component analysis, and T-test).

B. *Example 2 - DataAnalyticsWorkflow*

A PHPDM model should also specify the *Stakeholders* of the policy making process, i.e., the human actors who may participate in it. These participants of the process may express *Positions* over the different action options that are available in the process. A position expressed by a stakeholder can be a *SupportivePosition* (i.e., a position that supports the advocation of the action), an *OpposingPosition* (i.e., a position that is negative to the advocation of the action) or a *NeutralPosition* (i.e., a decision indicating that the stakeholder neither supports nor objects to the action). A stakeholder may express Supportive, Opposing or Neutral positions for one or more alternative actions but cannot express two different Positions for the same alternative.

VI. The Methodology

Figure 2 shows the user interactions with our tool. A policy maker interacts with a Dashboard in order to define a PHPDM model for a given policy using the PHPDM Language.[1] This activity is supported by a PHPDM specification service (PHPDM e-services) to ease the policy definition. Once the policy model is completely defined, the portion related to the *DataAnalyticsWorkflow* is used to trigger the analytics executions instrumenting the Big Data Engine (i.e., end-user decides when execution will be initiated). The *DataAnalyticsWorkflow* entity, as specified in the PHPDM model (Module 3) of Section IV is capable to detail any type of analytics workflow including, descriptive, predictive and prescriptive ones.

[1] We assume that the policy maker is assisted by a data scientist and a domain expert while designing the policy as is normally happening nowadays.

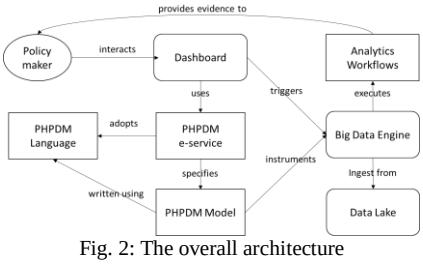

Fig. 2: The overall architecture

Analytics execution is not trivial since the *DataAnalytics-Workflow* is expressed as OWL and have to be transformed in a format that can be executed.

This transformation is similar to the transformation provided by a computer programming compiler that receives a program (i.e., the PHPDM model) in a specific framework (i.e., the PHPDM framework) made of instructions in a specific order and transform it into an executable for the specific target platform (i.e., the BDA). In case presented (e.g., example 2), the program is the instantiation of the Data Analytics Workflow, instructions are the Task (e.g., *DataProcessingTasks*), and the target platform is the Apache Spark-based BDE.

Definition 1 (PHPDM transformation): A PHPDM Data Analytics Workflow transformation F_C (L) is a function that receives as input a valid instance L of the PHPDM DataAnalyticsWorkflow and returns as output a valid Executable Big Data Workflow for a specific BDE. Specifically, it maps for each relevant task instances in L (e.g., chosen method filtering, joining, statistical analysis, time restrictions) a BDE capabilities c in C, where C is named as capability catalogue. The capability mapping is such that the produced output of each selected capability is compatible with the *OutputDataSpecification* of the PHPDM *DataAnalyticsWorkflow*. It also transforms the OWL workflow description into a Workflow script orchestrating the capabilities in a way that it is executable by the target BDA

The analytics capabilities catalogue C of our BDA is designed to match the *Operation*, *Algorithm* and the various type of *Tasks* entities that constitutes a given *DataAnalytics-Workflow*.

After the transformation, the BDA is then used to execute the analytics on a given data lake and generating evidence for the policy maker. Note that the structure/format of the data to be used for the analytics must be known when the PHPDM model is generated. In case the structure is not compatible with the one required by a given capability c in C a specific data format transformation task is prepended in the *DataAnalyticsWorkflow* (e.g., just after the data ingestion) in order to grant the compatibility.

VII. Big Data Analytics Engine

Our BDA similarly to the one used in [24] leverages on Apache ecosystem. More in details, the analytics engine is based on the Apache Spark [48] big data processing framework. It runs on the cluster consisting of virtual machines that provide its computational power to run data analytical tasks. In order to exploit the advantage of distributed calculations, the cluster management system Hadoop YARN [49] is deployed. Such solution implements a robust mechanism to allocate the load among computational units. It worth to mention that MapReduce, usually associated with Hadoop is not a part of our big data paradigm. Instead, all tasks are processed by Spark. The choice is dictated by advantage in computational speed and disk space allocation efficiency [50]. Big data engine uses Spark MLlib[2] to perform machine learning and graph analysis. It allows processing different dataset transformations, feature extraction and selection. This big data library provides a rich choice of classification, regression, clustering and filtering algorithms. The BDA capabilities C are obtained wrapping the MLlib as well as other libraries and ad hoc defined algorithms. The wrapper is used to standardize the interface in order to be easily connected to a workflow.

Every entity in the PHPDM model has an executable counterpart in the BDA engine that requires to be orchestrated in an executable workflow. As workflow manager we adopted Oozie[3] that allows to define complex workflows and to schedule their executions.

C. Example 3 (OneWayAnova Workflow)

Let us consider a PHPDM DataAnalyticsWorkflow of Example 2 as input to the BDA. Note that for simplicity no data transformation is needed, and the data ingested by the Data Lake is in the needed format. The transformation function F_C maps the required OneWayAnova to the relative capability *c* of the BDE for executing the OneWayAnova as requested and capable to produce an output compliant to the ANOVASpec definition. After the mapping Fc transform the OWL workflow into the Oozie orchestration.

The Oozie scheduling capabilities used for instance in the framework of refreshing a given Machine Learning model offline in order to perform online predictions.

Our BDA engine is focused to handle batch processing giving the nature of the policy making process, but it is capable to handle micro batches via spark stream if needed.

Prior to the workflow execution, BDA engine loads the input datasets from the data lake using predefined ingestion procedures. For the sake of simplicity among the other possibilities in this paper we consider, as landing platform for ingestion the HBase distributed database, which belongs to the Hadoop ecosystem. Such fact makes it suitable for mapping its tables as an input to the Spark jobs.

The outcome of the Spark jobs execution is combined by BDA engine and later is returned to the policy maker via a Dashboard (onwards PHPDM e-service). Therefore, the policy making perspective of the task execution is simplified to a single click of a button resulting in a numerical or graphical representation of the output supporting the decision-making process.

VIII. PHPDM e-Service

The PHPDM Specification Tool (PHPDM e-service) is the component instantiates the aforementioned ontology framework and allows end-users to administer decision models and their execution. This service assists them in defining suitable instances of PHPDM models, in accordance with a predefined

[2] https://spark.apache.org/mllib/

[3] http://oozie.apache.org/

template of a particular model. Appropriate functions guide the end-user in defining those models by dynamically adapting the possible choices (e.g., of input datasets and parameters, of method to be applied upon them, of thresholds or other execution criteria to be fulfilled) logically defined by the ontology.

When a model instance is completed (i.e., all necessary input parameters are defined) and validated (i.e., conform to the specification), the end-user can commit it through the PHPDM e-service for execution. In such case, the execution action triggers necessary data transformations and packaging them with configuration settings necessary for the execution by the BDA engine (i.e., input parameters, method related parameters, pointers to big data structures, execution criteria), for transforming the model specification into an executable workflow for the BDA Engine, and then of invoking the BDA Engine.

In accordance to the specification:

- Each policy model is aimed at one Goal;
- The Goal has a description and a rationale and is refined into multiple Objectives;
- Each Objective has a description and a rationale and can be addressed by one or more Policy Actions;
  - A Policy Action can be alternative, dependent, or prerequisite to another policy action;
- Each Policy Action is correlated to one data analytics Workflow;
- The data analytics Workflow is composed of one or many data analytics Tasks;
- A data analytics Task can be a social media analytics task, a simulation task, a statistical analysis task, a data processing task, a text mining task or a data mining task;
  - Each Task utilizes a method, which according to the type is an operation (for data processing tasks) or an algorithm (a data mining task utilizes a data mining algorithm, a statistical analysis task utilizes a statistical analysis algorithm, a text mining task utilizes a text mining algorithm, etc.);
  - Each task also has one or many input datasets and one or many output datasets;
  - Each dataset has a data specification.

IX. PHPDM Use Case

PHP makers are expected to make use of the part of the framework that enables the specification of policy goals and objectives (Module 1), the stakeholders who may be involved in the decision-making process, and the potential policy actions (Module 2). Clinicians are expected to make use of the part of the framework that enables the specification of policy objectives, potential policy actions, and the evidence and criteria required for making decisions regarding the actions (Module 1). They may also be involved in the identification of the data sets and the analytic processes that need to be analysed for generating the evidence (Module 3). Data scientists make use of the part of the framework that enables the specification of the data sets and the analytic processes that need to be analysed for generating the evidence (Module 3), and the criteria for establishing the plausibility of the generated evidence in support of different actions (Module 1).

During the formation of a PHPDM model to aid decision making, the overall goal of the model would be set by policy making authorities. The objectives of the model, (i.e., to introduce interventions to address prevention of HA usage due to occupation, due to education level, or due to age) could be set following a dialogue between clinicians and policy makers that identifies the particular factors as worth exploring further before making decisions on the relevant interventions.

The specific data analytic procedures that will be used to explore such factors would typically be identified through a dialogue between clinicians and data scientists, through which the most appropriate analytic methods are established following consideration of the types of the input data involved (e.g., numeric vs. nominal data), and the conditions that should be satisfied by the available data set in order for an analytics technique to be expected to produce meaningful results. Various forms of statistical analysis, such as linear regression, could for example be deemed as non-appropriate if the independent variables are themselves linearly independent (i.e., it is not possible to predict any of them through a linear combination of the other).

Data scientists provide the policy authority representatives and the representative of the clinicians or drawn from other organisations with established expertise on the subject.

In the following, focused is given on the Data processing part of the above architecture. As such, we present a use case of a EHP in question – in the context of EVOTION data - that utilizes some basic features (e.g., basic statistical methods, filtering mechanism). Investigate whether self-management of hearing health interventions on a population scale have positive effects on wellbeing and quality of life. Notably, workflows and methods applied for the purpose of the following scenario, they are not the only ones applicable, (all MLIB[4] methods supported by the implemented BDA engine [31]).

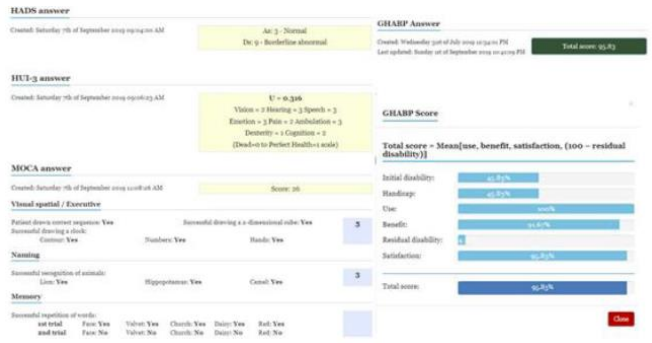

Fig. 3: Output (score) of each questionnaire supported by the EVOTION platform.

This Hearing Loss related use case considers data analytics for studying whether the Hearing Aid devices and the Auditory Training (AT) rehabilitation services provided during the EVOTION study for a period of one year (divided into 2 semi-annual periods) produced positive (or negative) effects on hearing-aid (HA) use and on mitigating cognitive and auditory process deterioration based on (Figure 3): (i) Montreal Cognitive

[4] APACHE SPARK Mlib: https://spark.apache.org/mllib/

Assessment (MoCA) total score (1st vs last visit); (ii) Hospital Anxiety and Depression Scale (HADS) anxiety score (1st vs last visit); (iii) Glasgow Hearing Aid Benefit Profile (GHABP) total score; and (iv) Monthly Total HA-Usage derived during the specific period. This is to support policy makers in the definition of actions targeted to determine whether the use of HA in conjunction to the use of various mobile tests performed (self-management hearing health interventions) helped EVOTION participants in utilizing more effectively the HA devices and had positive effects on their wellbeing and quality of life. Thus, the Goal of the PHPDM is improving HA effectiveness.

An example of a EVOTION Policy definition and execution is the following[5]: Investigate whether self-management of hearing health interventions on a population scale have positive effects on wellbeing and quality of life.

Definition

- Choose "active" patients for a period of one year (two half-yearly periods A and B) (Active patients to be considered those having more than 3 hours of total monthly HA-Usage for at least one month during the specified period);
  - Disregard those having MOCA total score outliers (MOCA records having total score < 17);
- Compare total HA usage when use is at its peak (study period: Sep 2018 - Aug 2019);
- Parameters to be compared:
  - Average MOCA total score (1st vs last visit);
  - Average HADS anxiety score (1st vs last visit);
  - Average GHABP total score (score > 60);
  - Average monthly Total HA-Usage (period A vs period B);
- Update, re-validate and re-execute the policy;

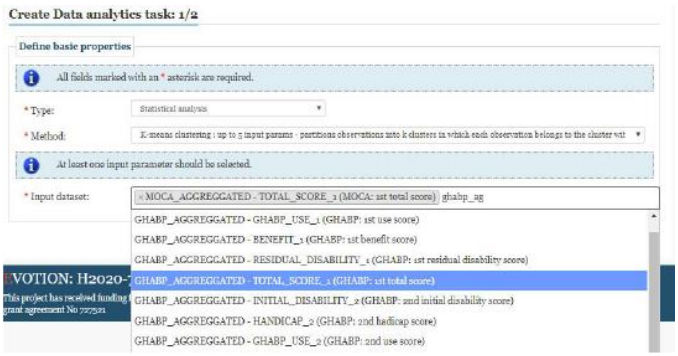

Fig. 4: Example of Task creation (1/2): K-means clustering on 2 parameters.

Fig. 5: Example of Task creation (2/2): Define K (number of groups).

Sequence of actions

- Create Workflows (as in Figures 4 and 5) to:
  - Select all questionaries' answers parameters to be compared and disregard records having null values
  - Calculate mean and median of MOCA scores (1st vs Last visit)
  - Calculate mean and median of HADS scores -(1st vs Last visit)
  - Calculate mean and median of GHABP scores (Part A, Part B)

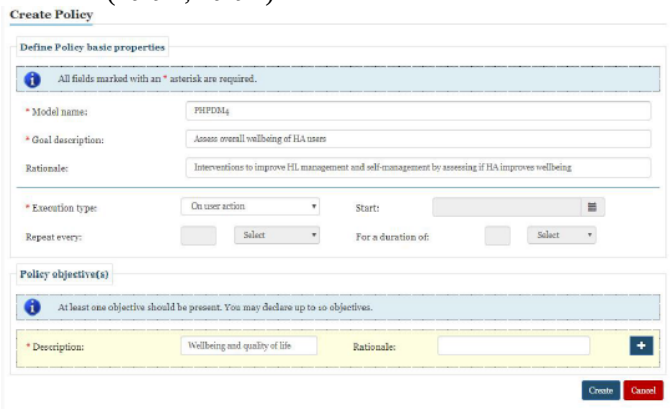

Fig. 6: Example of Policy creation.

- Create a Policy (as in Figure 6)

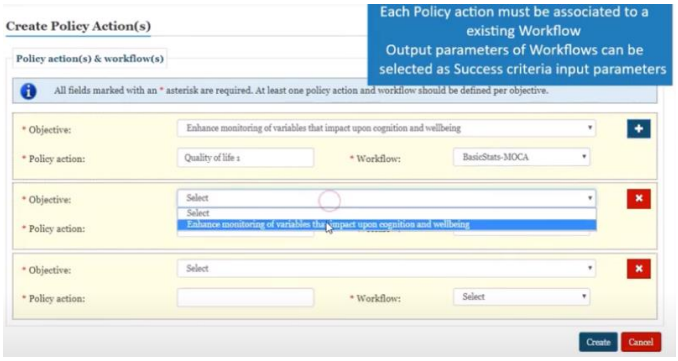

Fig. 7: Example of Policy Criteria creation.

  - Create 3 Policy actions to be associated to previously executed Workflows (as in Figure 7)

[5] A step-by-step creation and execution demonstrator is available online at: https://www.youtube.com/watch?v=sm_4pUEpYLI

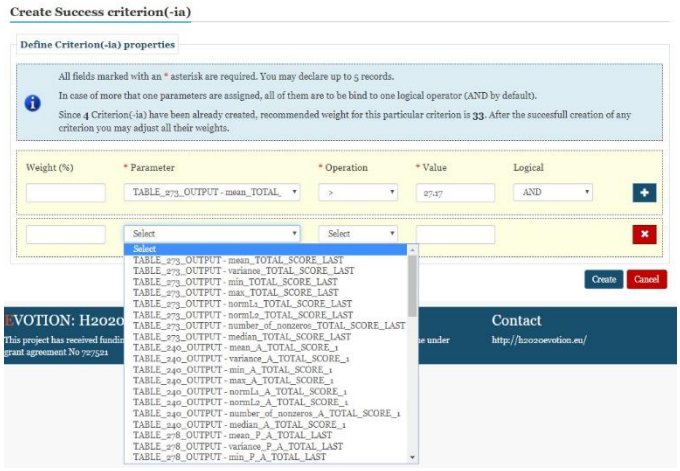

Fig. 8: Example of Policy Criteria creation.

- Create 3 Success criteria, (related to the previously defined Policy actions (as in Figure 8)

- Create Workflows to:
  - Filter-HA-Usage: select records from Sep 1, 2018 to Aug 31, 2019
  - Disregard periods of inactivity (records having Total monthly HA-Usage < 1800 minutes)
  - Split-HA-Usage-2Periods: spit dataset in 2 periods A and B (Period A: Sep 1, 2018 to Feb 28, 2019, Period B: Mar 1, 2019 to Aug 31, 2019)
  - BasicStats-HA-Usage-2Periods: calculate means and averages for those periods

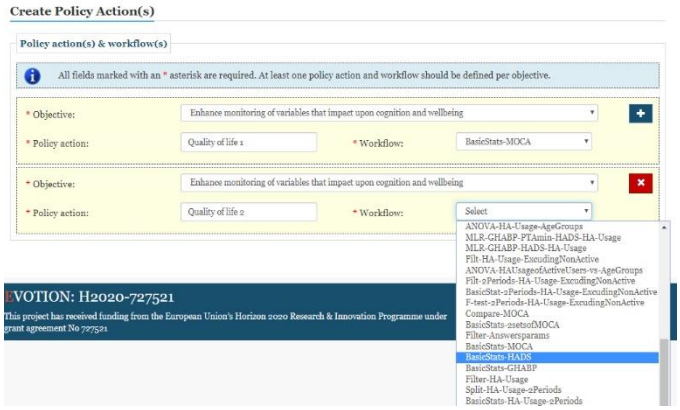

Fig. 9: Example of Policy Criteria - Workflows associations.

- Associate Workflows output parameters with Success Criteria (as in Figure 9);
- Execute policy (as in Figure 10).

The application of big data to support evidence-based public health policy decision-making for hearing is discussed in [51].

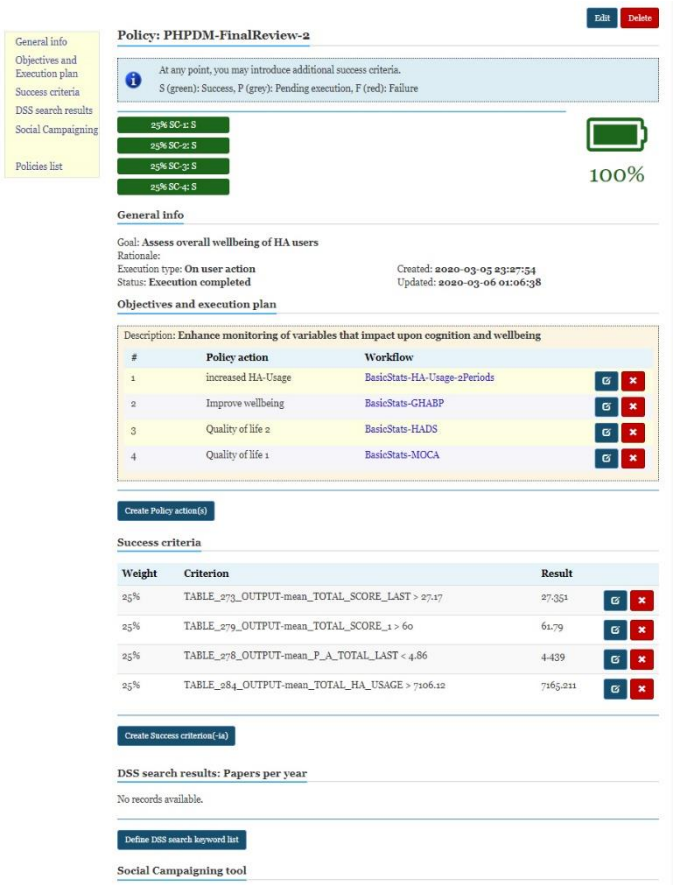

Fig. 10: Example of Policy execution output: Output of a successful (all execution criteria have been met) PHPDM policy.

X. Performance Evaluation

In this section we present a preliminary performance evaluation on our BDE. To this purpose, we describe the testing BDE environment and the dataset, implementations as Java Workflows, and finally the performance of the Analytics Workflows varying the dataset dimensions is evaluated. In the following we describe the implementation of the specific transformation for the example in section VIII and presents some preliminary performance evaluation.

A. Testing environment

For the evaluation of a basic PHPDM model that entails the execution of a performance-intensive big data analytics task (an ANOVA task), we consider a dataset of synthetically generated data made by 10M of records. Experiments have been performed on a DELL VRTX Blade server made by two blade PowerEdge M630 equipped with 2 Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz and 192GB of RAM. On top of this blade server, an OpenStack instance was mounted over the BDE based on 5 nodes.

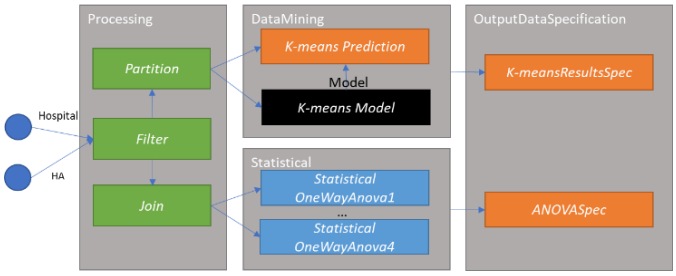

Fig 11: The Workflows as implemented in BDA.

Figure 11 shows the structure of the implementation of this primitive model. The execution of which entails that statistical and data mining tasks are executed in parallel as well as internal Statistical evaluation are made by four parallel OneWayAnova analytic Task. We also note that for the shake of this experiments we adopted K equal to 3 for the K-means clustering.

In this example we also consider a partitioning processing Task focused on select the portion of data to be used for building the model for the clustering compared with the ones that are used to predict the cluster membership. K-means can be also used as unique providing prediction while modelling the clusters. We adopt Spark MLlib k-means on one side while we implemented OneWayAnova from scratch using spark on the other.

B. *Performance Evaluation Results*

We evaluated the performance of our approach executing the above WF1 varying the dimension of the dataset. We evaluate the computational time required for i) each pre-processing step, ii) analytics steps. Figure 8 shows the performance results in terms of average computational time evaluated at the average of 10 executions.

Processing and analytics (statistical and data mining together) execution time are presented increasing the number of records in the dataset. The processing time are less impacted by the number of records compared to the analytics processing. In total for processing 10M of records executing the WF1 the BDA requires almost 500 seconds (Figure 12).

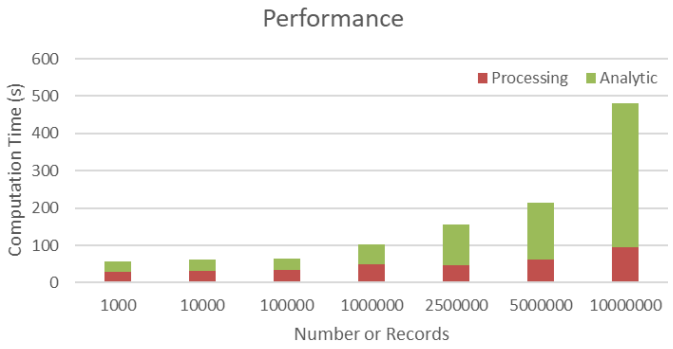

Fig 12: Performance evaluation of WF1 increasing the number of records.

During the utility analysis conducted to evaluate the utility and the usability of the EVOTION platform to the policy makers, almost all respondents (97.4%) consider the e-service as a useful one and recognized its usefulness. In order to demonstrate an early version of the functionality to relevant stakeholders, a series of workshops conducted (London UK, Osijek Croatia, Sofia Bulgaria, and Warsaw Poland) finding of which reported in [30]. Overall, stakeholders consider the EVOTION platform a useful framework, and expressed an interest to utilise it to generate evidence-based, high-quality policy recommendations. Still, late-stage usability evaluation is scheduled to be conducted to determine on how well end-users can learn and use the service to achieve their goals.

XI. Comparison with Existing Data Analytics Technologies/Tools

A. *The R Project for Statistical Computing*

R [52] is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language [53] and environment which was developed at Bell Laboratories. R can be considered as a different implementation of S. S is a language for the manipulation of objects. It aims to be both an interactive language (like, for example, a Unix shell language) as well as a complete programming language with some convenient object-oriented features.

Although R offers a JAVA API and is extensible, it is rather complicated to work with R and big data.

B. *RapidMiner*

RapidMiner [54] is an open-source software platform for data science teams that unites data preparation, machine learning, and predictive model deployment. It is written in JAVA programming language.

Although it is relatively easy to run analysis in Rapid Miner GUI and it also offers a JAVA API, it is not fit for our purpose, as it is not to be used for big data analytics.

C. *Orange*

Orange [55] is an open-source software suite for machine learning & data mining. It best aids the data visualization and is a component-based software. It has been written in Python computing language.

Although Orange is easy to run analytics through its GUI, it does not offer an API, so it is not fit for our purpose.

D. *KNIME*

KNIME [56] is an open-source integration platform for data analytics and reporting. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together.

Although KNIME offers an easy-to-use interface for running analytics with very useful reporting and visualization capabilities, it is not to be used for big data analytics.

E. *Apache Mahout*

Apache Mahout [57] is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering.

Although Apache mahout is fit for our purpose, as it is built to work with big data and it is written in JAVA, it does not have any user interface to run data analytics with.

F. Weka

Weka [58], [59] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own JAVA code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

G. Comparison

Below in we summarize and compare the features of the data analytics technologies/tools reviewed above. We use the following criteria for the comparison of the data analytics technologies: whether they incorporate a graphical user interface (GUI), whether they include a JAVA API, whether they are interoperable with big data analytics technologies (Big Data) and whether they enable the user to specify workflows by programmatically (scripting) or with a use of a GUI (Graphics).

TABLE 1
DATA ANALYTICS TECHNOLOGIES/TOOLS FEATURES

Data analytics technology/tool	GUI	JAVA API	Big Data	Workflows
R	X	✓	X	Scripting
RapidMiner	✓	✓	X	Graphics
Orange	✓	X	X	Graphics
KNIME	✓	✓	X	Graphics
Apache Mahout	X	✓	✓	Scripting
Weka	✓	✓	✓	Graphics

SUMMARY

While the use of big data analytics in healthcare for policy-making is still in its infancy, the need among those health professionals for support in monitoring policy's implementation by applying a reliable collection of indicators measuring the day-to-day activities is growing. As documented in the case of the recent pandemic, support of evidence-based analysis is considered a myth-busting factor and as such user-friendly big data analysis infrastructures serve this dual purpose. The research prototype EVOTION platform illustrates an example on how EBPHP can be supported, to preform analytics in a well-defined way that focuses on reducing the workload of the end-users participating in the formulation, execution and analysis of a polity, as well as to acquire high performance in big data processing.

XII. ETHICAL ISSUES USING THE PLATFORM TO COLLECT HL DATA AND PRIVACY CONSIDERATIONS

Prior to starting the recruitment of patient and the collection of prospective and retrospective data into the EVOTION platform, the EVOTION consortium obtained ethics approval. The process and the documentation for this, including consent forms, have been included in Study Protocol of EVOTION. Consent forms have been formulated to be fully compliant with the specifications of the General Data Protection Regulation (GDPR). The EVOTION platform, by virtue of its design, supports privacy. For stored data, personally identifiable information (PII) for the subject of the data is masked or removed from it altogether (GDPR compliance presented in [60]).

EVOTION dataset is approved by several Ethics approval protocols and includes: a) Retrospective data: patients demographics, HL levels, cause and duration of HL, medical history and HA usage data, Audiograms, b) Prospective data: audiological and other assessments (Montreal Cognitive Assessment, Pure Tone Audiometry, Hospital Anxiety and Depression Scale, Glasgow Hearing Aid Benefit and Health Utility Index Mark-3, HA: Hearing Aid, REM: Real Ear Measurement).

ACKNOWLEDGMENT

This project has received funding from the European Commission's Horizon 2020 research and innovation program under grant agreement no. 727521.

REFERENCES

[1] World Health Organization, "Key Policy Issues in Long-Term Care," *World Heal. Organ.*, pp. 139–190, 2003, [Online]. Available: http://www.who.int/chp/knowledge/publications/policy_issues_ltc.pdf.

[2] K. O'Neill, K. Viswanathan, E. Celades, and T. Boerma, "Monitoring, evaluation and review of national health strategies," in *Strategizing national health in the 21st century: a handbook*, 2016.

[3] WHO, "Health policy." http://www.who.int/topics/health_policy/en/ (accessed Sep. 10, 2018).

[4] R. C. Brownson, J. F. Chriqui, and K. A. Stamatakis, "Understanding evidence-based public health policy," *Am. J. Public Health*, vol. 99, no. 9, pp. 1576–1583, 2009, doi: 10.2105/AJPH.2008.156224.

[5] F. Rajabi, "Evidence-informed health policy making: The role of policy brief," *International Journal of Preventive Medicine*. 2012.

[6] B. Hawkins and J. Parkhurst, "The 'good governance' of evidence in health policy," *Evid. Policy*, 2016, doi: 10.1332/174426415X14430058455412.

[7] J. V. Gautam, H. B. Prajapati, V. K. Dabhi, and S. Chaudhary, "A survey on job scheduling algorithms in Big data processing," 2015, doi: 10.1109/ICECCT.2015.7226035.

[8] B. Ye, I. Basdekis, M. Smyrlis, G. Spanoudakis, and K. Koloutsou, "A big data repository and architecture for managing hearing loss related data," 2018, doi: 10.1109/BHI.2018.8333397.

[9] M. Marjani *et al.*, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017, doi: 10.1109/ACCESS.2017.2689040.

[10] M. Viceconti, P. Hunter, and R. Hose, "Big Data, Big Knowledge: Big Data for Personalized Healthcare," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1209–1215, 2015, doi: 10.1109/JBHI.2015.2406883.

[11] M. Prasinos, G. Spanoudakis, and D. Koutsouris, "Towards a model-driven platform for evidence based public health policy making," 2017, doi: 10.18293/SEKE2017-180.

[12] G. Spanoudakis, P. Katrakazas, D. Koutsouris, D. Kikidis, A. Bibas,

and N. H. Pontopidan, “Public Health Policy for Management of Hearing Impairments Based on Big Data Analytics: EVOTION at Genesis,” *2017 IEEE 17th Int. Conf. Bioinforma. Bioeng.*, pp. 525–530, 2017, doi: 10.1109/BIBE.2017.00006.

[13] World Health Organization, “Ear and hearing care: situation analysis tool.” 2015, [Online]. Available: https://apps.who.int/iris/handle/10665/206141.

[14] E. A. Akl *et al.*, “The SPARK Tool to prioritise questions for systematic reviews in health policy and systems research: Development and initial validation,” *Heal. Res. Policy Syst.*, vol. 15, no. 1, pp. 1–7, 2017, doi: 10.1186/s12961-017-0242-4.

[15] S. R. Makkar, F. Gilham, A. Williamson, and K. Bisset, “Usage of an online tool to help policymakers better engage with research: Web CIPHER,” *Implement. Sci.*, vol. 10, no. 1, pp. 1–11, 2015, doi: 10.1186/s13012-015-0241-1.

[16] National Collaborating Centre for Methods and Tools, “NCCMT Webinar: Applicability and Transferability of Evidence (A&T) Tool,” 2016. https://www.youtube.com/watch?v=Fk28OMZMFlM.

[17] Deloitte Access Economics, “Evaluation framework for adult hearing services in England,” *Deloitte Access Econ.*, no. January, 2013.

[18] Canadian Foundation of Healthcare Improvement, “Accelerating Healthcare Improvement : Canadian Foundation for Healthcare Improvement ’ s Assessment Tool (CFHI Assessment Tool ™),” 2014, [Online]. Available: http://www.cfhi-fcass.ca/Libraries/Documents/Self-Assessment-Tool-2014-E.sflb.ashx.

[19] J. A. Jacobs, E. Jones, B. A. Gabella, B. Spring, and R. C. Brownson, “Tools for implementing an evidence-based approach in public health practice,” *Prev. Chronic Dis.*, 2012, doi: 10.5888/pcd9.110324.

[20] R. Wigglesworth, “Can big data revolutionise policymaking by governments,” *Financ. Times*, vol. 31, 2018.

[21] B. Holford, “How big data is helping to transform the defence sector,” *Public Technology Net*, 2017.

[22] D. Broeders, E. Schrijvers, B. van der Sloot, R. van Brakel, J. de Hoog, and E. Hirsch Ballin, “Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data,” *Comput. Law Secur. Rev.*, 2017, doi: 10.1016/j.clsr.2017.03.002.

[23] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014, doi: 10.1186/2047-2501-2-3.

[24] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, “Health-CPS: Healthcare cyber-physical system assisted by cloud and big data,” *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, 2017, doi: 10.1109/JSYST.2015.2460747.

[25] M. M. Rathore, A. Paul, A. Ahmad, M. Anisetti, and G. Jeon, “Hadoop-Based Intelligent Care System (HICS),” *ACM Trans. Internet Technol.*, vol. 18, no. 1, pp. 1–24, 2017, doi: 10.1145/3108936.

[26] B. Fabian, T. Ermakova, and P. Junghanns, “Collaborative and secure sharing of healthcare data in multi-clouds,” *Inf. Syst.*, vol. 48, pp. 132–150, 2015, doi: 10.1016/j.is.2014.05.004.

[27] D. Brdari, “A Data-informed Public Health Policy-Makers Platform,” 2020.

[28] K. Doka, N. Papailiou, D. Tsoumakos, C. Mantas, and N. Koziris, “IReS: Intelligent, multi-engine Resource Scheduler for big data analytics workflows,” 2015, doi: 10.1145/2723372.2735377.

[29] “The EVOTION project: Preventing deafness and hearing loss,” 2019. https://www.openaccessgovernment.org/the-evotion-project-preventing-deafness-and-hearing-loss/79679/.

[30] G. Dritsakis *et al.*, “Public health policy-making for hearing loss: stakeholders’ evaluation of a novel eHealth tool,” *Heal. Res. Policy Syst.*, vol. 18, no. 1, 2020, doi: 10.1186/s12961-020-00637-2.

[31] M. Anisetti, C. Ardagna, V. Bellandi, M. Cremonini, F. Frati, and E. Damiani, “Privacy-aware Big Data Analytics as a Service for Public Health Policies in Smart Cities,” *Sustain. Cities Soc.*, 2018.

[32] E. N. Loukis, “An ontology for G2G collaboration in public policy making, implementation and evaluation,” *Artif. Intell. Law*, vol. 15, no. 1, pp. 19–48, 2007, doi: 10.1007/s10506-007-9041-5.

[33] P. Panov, L. Soldatova, and S. Džeroski, *Ontology of core data mining entities*, vol. 28, no. 5–6. 2014.

[34] B. Smith and P. Grenon, “Basic formal ontology (bfo),” *INFOMIS Reports*, 2006.

[35] B. Smith *et al.*, “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration,” *Nat. Biotechnol.*, vol. 25, p. 1251, Nov. 2007.

[36] A. Bandrowski *et al.*, “The Ontology for Biomedical Investigations,” *PLoS One*, 2016, doi: 10.1371/journal.pone.0154556.

[37] W. Ceusters, “An information artifact ontology perspective on data collections and associated representational artifacts,” 2012, doi: 10.3233/978-1-61499-101-4-68.

[38] P. Panov, “A MODULAR ONTOLOGY OF DATA MINING,” *Dr. Diss.*, 2002.

[39] “STATO Ontology.” http://stato-ontology.org.

[40] W3C, “OWL W3C Website,” 2012. https://www.w3.org/OWL/ (accessed Nov. 08, 2017).

[41] Stanford Center for Biomedical Informatics Research, “Protégé Website,” 2016. http://protege.stanford.edu.

[42] W3C, “RDF W3C Website,” 2014. https://www.w3.org/RDF/.

[43] W3C, “Manchester OWL Syntax W3C Website,” 2012. https://www.w3.org/TR/owl2-manchester-syntax/.

[44] W3C, “Turtle OWL Syntax W3C Website,” 2014. https://www.w3.org/TR/turtle/.

[45] W3C, “RDF XML Syntax W3C Website,” 2014. https://www.w3.org/TR/rdf-syntax-grammar/.

[46] W3C, “OWL2/XML Syntax W3C Website,” 2012. https://www.w3.org/TR/2012/REC-owl2-xml-serialization-20121211/.

[47] OMG, “Unified Modelling Language (UML), Version 2.5,” 2015. http://www.omg.org/spec/UML/2.5/PDF (accessed Nov. 08, 2017).

[48] M. Zaharia *et al.*, “Apache Spark: a unified engine for big data processing,” *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016, doi: 10.1145/2934664.

[49] V. K. Vavilapalli *et al.*, “Apache hadoop YARN: Yet another resource negotiator,” 2013, doi: 10.1145/2523616.2523633.

[50] P. J. Verma and A. Patel, “Comparison of MapReduce and Spark Programming Frameworks for Big Data Analytics on HDFS,” *Ijcsc*, vol. 7, no. 2, pp. 80–84, 2016, doi: 10.090592/IJCSC.2016.113.

[51] G. H. Saunders *et al.*, “Application of Big Data to Support Evidence-Based Public Health Policy Decision-Making for Hearing,” *Ear Hear.*, 2020, doi: 10.1097/AUD.0000000000000850.

[52] R Development Core Team, “R: A Language and Environment for Statistical Computing,” *R Found. Stat. Comput. Vienna Austria*, 2016, doi: 10.1038/sj.hdy.6800737.

[53] J. M. Chambers, *Programming with data: A guide to the S language*. Springer Science & Business Media, 1998.

[54] M. Hofmann; and R. Klinkenberg;, “RapidMiner: Data Mining Use Cases and Business Analytics Applications,” *Zhurnal Eksp. i Teor. Fiz.*, 2013, doi: 78-1-4822-0550-3.

[55] J. Demšar *et al.*, “Orange: Data Mining Toolbox in Python,” *J. Mach. Learn. Res.*, 2013.

[56] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, and B. Meinl T.... & Wiswedel, “KNIME-the Konstanz information miner: version 2.0 and beyond.,” *AcM SIGKDD Explor. Newsl.*, 2009.

[57] T. A. S. Foundation, “Apache Mahout: Scalable machine learning and data mining,” *Apache Mahout*, 2016. .

[58] and I. H. W. Eibe Frank, Mark A. Hall, “The WEKA Workbench. Online Appendix,” in *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edi., Morgan Kaufmann, 2016.

[59] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009, doi: 10.1145/1656274.1656278.

[60] I. Basdekis, K. Pozdniakov, M. Prasinos, and K. Koloutsou, “Evidence based public health policy making: Tool support,” 2019, doi: 10.1109/SERVICES.2019.00080.