Automatized bioinformatics data integration in a Hadoop-based data lake

Júlia Colleoni Couto, Olimar Teixeira Borges, and Duncan Dubugras Ruiz

School of Technology, PUCRS University,

Abstract. When we work in a data lake, data integration is not easy, mainly because the data is usually stored in raw format. Manually performing data integration is a time-consuming task that requires the supervision of a specialist, which can make mistakes or not be able to see the optimal point for data integration among two or more datasets. This paper presents a model to perform heterogeneous in-memory data integration in a Hadoop-based data lake based on a top-k set similarity approach. Our main contribution is the process of ingesting, storing, processing, integrating, and visualizing the data integration points. The algorithm for data integration is based on the Overlap coefficient since it presented better results when compared with the set similarity metrics Jaccard, Sørensen-Dice, and the Tversky index. We tested our model applying it on eight bioinformatics-domain datasets. Our model presents better results when compared to an analysis of a specialist, and we expect our model can be reused for other domains of datasets.

Keywords: Data integration, Data lake, Apache Hadoop, Bioinformatics.

1 Introduction

Data integration is a challenging task, even more nowadays where we deal with the V's for big data, such as variety, variability, and volume (Searls [1]; Lin et al.[2]; Alserafi et al. [3]). Regarding the variety of data to be integrated into data lakes, having different types of data can be considered one of the most difficult challenges, even more because most datasets may contain unstructured or semi-structured information (Dabbèchi et al. [4]). According to Hai, Quix, and Zhou [5], it is very onerous to perform interesting integrative queries over distinct types of datasets. Another challenge is the high-dimensionality data that may be stored in the data lake. To compute the similarity for that high-dimensional data is expensive. Checking whether the tables are joinable or not is time-consuming because of the large number of tables that may have in a data lake (Dong et al. [6])

To manually analyze different datasets for data integration, a person must check the attributes and at least a dataset sample. To perform a more elaborated work, the person must look for the data dictionary of each dataset, and sometimes it is not easily available. According to Sawadogo and Darmont [7], it is a problem since it is time-consuming, error-prone, and can lead to data inconsistency. Among the methods for data integration, the logic-based ones that consider the dataframes as sets, such as the based on the overlap of the values, could provide useful solutions (Levy [8]).

This paper presents a model we developed to perform heterogeneous data integration, taking advantage of a data lake we build based on Hadoop. The integration model is based on schema matching techniques, such as row content-based overlapping. To do so, we first define the datasets for the experiments, related to the domain of bioinformatics. Then, we build a data lake to ingest, store, process, and visualize the data. We use Apache Nifi for data ingestion and the HDFS (Hadoop Distributed File System) for data storage. We process the data using Python, and we create visualizations of the data using Neo4J.

David C. Wyld et al. (Eds): AIAPP, NLPML, DMA, CRIS, SEC, CoSIT, SIGL - 2022
pp. 137-153, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120912

Our main contribution is a model that allows to quickly ingest different kinds of textual datasets, transform them into dataframes, and, using an approach based on in-memory set similarity for data integration, we suggest the top-k points of integration for the data. We present experiments with eight bioinformatics datasets, and we compare our approach with manual data integration performed by a domain specialist. Our paper can also be used as a guide to building a data lake from scratch.

In what follows, we investigate automatized data integration. Section 2 presents an explanation about the main topics that are essential for our study and related work. Section 4 describes the methodology we followed, and Section 5 presents the results we achieved. Section 6 discusses our results, challenges, and presents an example with the usefulness of our model. Finally, Section 7 summarizes our conclusions and future work.

2 Background

In this section, we briefly present the basic concepts related to our study. We summarize data integration, data lake, and present the datasets we used in our experiments. We finish by discussing the related work.

2.1 Data integration

Data integration deals with the problem of combining different data sources to provide the user with a unified view (Lenzerini [9]). There are different approaches for data integration, and we base our work on the top-k overlap set similarity problem: For all the attributes in all the dataframes, find the top fits for data integration, according to the intersection among the attributes' distinct values (Zhu et al. [10]). We based our work on an in-memory set similarity approach since the integration is executed using Python notebooks. As similarity metrics, we use the most well-known distance measures for sets similarity according to Ontanón (2020) [11]: Tverski's (Tversky [12]), Sørensen's index (Sørensen [13]), and Jaccard (Jaccard [14]), compared to the Szymkiewicz-Simpson overlap coefficient (Vijaymeena and Kavitha [15]). We develop our data integration based on a data lake.

2.2 Data lake

In a previous work (Couto et al. [16]), we define a *data lake* as a central repository for raw data storage, processing, and analysis, that can be used for unstructured, semi-structured, and structured datasets. A data lake can be composed of different software with its own tasks in an integrated ecosystem. It means we can have different software for data ingestion, storage, processing, presentation, and security, and they have to work together. The most used tool to create a data lake is Apache Hadoop [16]. Forster [17] states that Hadoop is the most used distributed platform for storage, processing, and analysis of big data. Hadoop is an Open-Source Software (OSS) developed in Java and maintained by the Apache Software Foundation [18]. Hadoop is based on the Google MapReduce paradigm and in Google File System, and it is mainly used for distributed processing in computer clusters. We populate our data lake with bioinformatics datasets.

2.3 Bioinformatics datasets

Bioinformatics is the product of the union of computer science and biology (Lesk [19]), where we use software to make inferences about datasets of modern molecular biology, so we can connect the data and extract valuable predictions. There are a lot of bioinformatics

datasets available, having the most variate information, formats, types, and size. Our study selected eight datasets to populate our data lake and work on automatized data integration. Table 1 presents the characteristics of each dataset, ordered by size from the smaller (DRUGBANK) to the larger (IID). Table 1 shows that we selected heterogeneous datasets, having varied sizes (from 1 MB to 1,8GB), from 13k entries to almost 1 million entries, with the number of attributes varying from 6 to 253. The datasets are also presented in different formats, such as TXT, XML, TSV, JSON, and via API.

Table 1. Characteristics of the bioinformatics datasets

Dataset	Size (MB)	Entries	Attributes	Format
DRUGBANK	0,95	13580	9	XML
DRUGBANK PROTEIN	1,40	26965	7	XML
OMIM	1,80	17092	14	TXT
DRUGCENTRAL	2,50	17390	19	TSV
MONDO	4,00	43233	12	JSON
DISGENET	10,30	84037	16	TSV
UNIPROT	30,20	565255	7	API
REACTOME	37,90	826877	6	TXT
IID	1800,00	975877	253	TXT

- OMIM (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2021 [20]): Online Mendelian Inheritance in Man - human genes and genetic phenotypes. We used the *genemap2* dataset.
- DISGENET (Pinero et al. [21]): Collections of genes and variants associated with human diseases.
- REACTOME (Jassal et al. [22]): We are using the *UniProt2Reactome* dataset. It is composed of reactions, proteins, and pathways.
- MONDO (Mungall et al. [23]): Ontology for disease definitions.
- DRUGBANK (Wishart et al. [24]): Pharmaceutical knowledge base, we split it into two dataframes: DRUGBANK and DRUGBANK_PROTEIN.
- IID (Kotlyar et al. [25]): Integrated Interactions Database - database of detected and predicted protein-protein interactions. We used the human data annotated dataset.
- DRUGCENTRAL (Avram et al. [26]): Online drug compendium - we use the drug-target interaction dataset.
- UNIPROT (Consortium [27]): We are using the *reviewed Swiss-Prot XML* dataset, a non-redundant and manually annotated database containing protein sequences.

2.4 Related work

The work of Cockell et al. [28] describes Ondex, a Data Integration Platform. They represent the data as a graph, where the nodes present the concepts, and the edges present the relations. They developed parsers to import the data to OXL format. Then, they use mappers and transformers to join different types of datasets, remove nodes that are not connected and add information to the network. Finally, they manually traversed using Ondex to search for interesting examples of drug repositioning. They use the following datasets: DRUGBANK, UNIPROT, HPRD, KEGG, PFam, SymAtlas, G-Sesame, OpenBabel, and BLAST. They use the cross-references presented on Uniprot to include accession numbers from other linked datasets (e.g., ENSEMBL, GO, OMIM, PRINTS).

Sellis et al. [29] use web services and ontologies to create Semantic Web services, to integrate three biological databases: EMBL, MEDLINE, and Array Express. They answer the following query: "for a given Nucleotide Number (EMBL database), find all experiments (Array Express database) and all publications (MEDLINE) which have taken place." They use OWL-S - an ontology-based on Web Ontology Languages that describe web services.

In his work, Hendler [30] discusses themes related to data integration, discovery, linked data, and the combination of structured data and unstructured data, and the author presents some theoretical approaches to deal with issues that come from heterogeneous datasets integration. For instance: use of natural language processing, graph databases, alignment using a third dataset.

Petermann et al. [31] developed a model named Business Intelligence with Integrated Instance Graphs, which they use for graph-based data integration and analysis. Their model has three types of graphs (separate graph databases): one for Unified Metadata (UMG, where the nodes are the classes and the edges are the associations), one for Integrated Instance (IIG, where the nodes are data objects and the edges are the relationships) and the last one for Business Transactions (BTG). In their process, they first perform metadata acquisition and integration for the UMG, then instance integration to create the IIG, generation of BTGs, and graph analytics. They use Neo4J.

Bradshaw et al. [32] developed an automatic and semi-automatic semantic data integration approach, based on concept bags, for synonyms and non-synonymous concepts. Concept bags are similar to word bags used in data mining. They compute the similarity between data elements and medical terms. To check the similarity, they use the Jaccard algorithm. They convert text or named entities in concept codes and then compare it using a vector-based analysis method. They use the following datasets: UMLS (315 entries), REDCap (899649 entries), Medical terms (60 entries). They state that their method presents the same or better performance when compared to other approaches.

Zhu et al. [10] develop JOSIE: an algorithm for JOining Search using Intersection Estimation. They use inverted indexes (mapping from words to their location - for quick search in text files). They work with the join table search problem: for a column C in a table, find other tables in the data lake where the intersection between column and C is high. They use two data lakes: Open Data and WebTables. They compare their results with MergeList-D and ProbeSet-D.

Zhang and Ives [33] develop JUNEAU, an approach to support multiple table relatedness measures, such as augmenting training data, finding potential features to extract, clean data, and finding joinable or linkable tables. They use pruning, top-k, and approximation strategies to return the tables that are most related.

When we compare our work with JOSIE [10] and JUNEAU [33], which are the most related, the main difference is related to our algorithm performing the data integration of more than two dataframes and simultaneously outputting it to take advantage of the raw data in the data lake.

3 Problem Statement

As stated by Khalid and Zimányi [34], managing and querying a data lake is a difficult task, mainly because the data is heterogeneous, may have replicas or versions, have a considerable volume, and present quality issues. In this sense, data integration, which represents 80-90% of the challenges for data scientists (Abadi et al. [35]), is a fundamental task to enable querying a data lake. However, integrating heterogeneous data into data lakes is a complex task, mainly due to the variety of data types (Hendler [30], Alrehamy

and Walker [36] that can compose the data lake. If we only talk about textual data, there are countless extensions and possible formatting, such as: .txt, .docx, .csv, .xls, .xml, .json and so on. Furthermore, the analysis for the integration depends on experts, often data scientists, who need to spend time inspecting data profile information, such as the types of each attribute, a sample of that data, or studying the data dictionary - when the dictionary is available. Finally, data integration is essential for extracting a more holistic view and information from the data lake, enabling us to make simple to complex queries and add value to the information.

Therefore, for the problem of automatic data integration in data lakes, the input would a number of heterogeneous datasets and a threshold that limits the integration points of interest. The output would be the points of integration among the datasets, and the evaluation measures would be the ones based on an expert evaluation of the integration points.

Regarding the complexity of the problem, according to Alserafi et al. [3], the equation to calculate the total number of comparisons that needed to be performed to find the columns candidate to data integration is

$$comparisons = \left[d \times \frac{d-1}{2} \right] \times m^2 \tag{1}$$

where d represents the number of datasets, and m represents the average number of attributes for each dataset. Considering our datasets (previously presented in Table 1), we have 344 attributes in total, considering 9 dataframes, then $m = 38$. Thus, we would have to perform about 51984 comparisons among the attributes.

4 Methods

The purpose of this study is to create a model for automating the integration of datasets. To do so, we use similarity measures to check the possibility of data integration in a Hadoop-based data lake. We started by creating the *system architecture* for the data lake, based on Docker containers. Then, we worked on *data management*. Finally, we present the *algorithm* we developed.

4.1 System architecture

Our data lake is supported by an Ubuntu 20 64-bit Linux server, having the following configuration: 16GB RAM DDR3, Processor Intel® Core(R) I7-4790 CPU@3.6GHz x 8, 1TB disk capacity. The data lake is composed of ten Docker containers:

1. Apache Nifi: a framework used for data ingestion.
2. Python - Jupyter Notebook: a programming language and a web application to run Python code, used for data processing.
3. Neo4J: a graph database used to visualize the integration among the dataframes.
4. Hadoop Namenode: the master node in the HDFS architecture
5. Hadoop History Server: keeps the logs of all the jobs that ran on Hadoop.
6. Hadoop Resource Manager: contains the YARN (Yet Another Resource Negotiator), a service that manages the resources and schedules/monitors the jobs.
7. Hadoop Node Manager: launches and manages the containers on a node.
8. Hadoop Datanodes: Three containers (Datanode1, Datanode2, Datanode3). The worker's nodes in the HDFS architecture, where the data is stored.

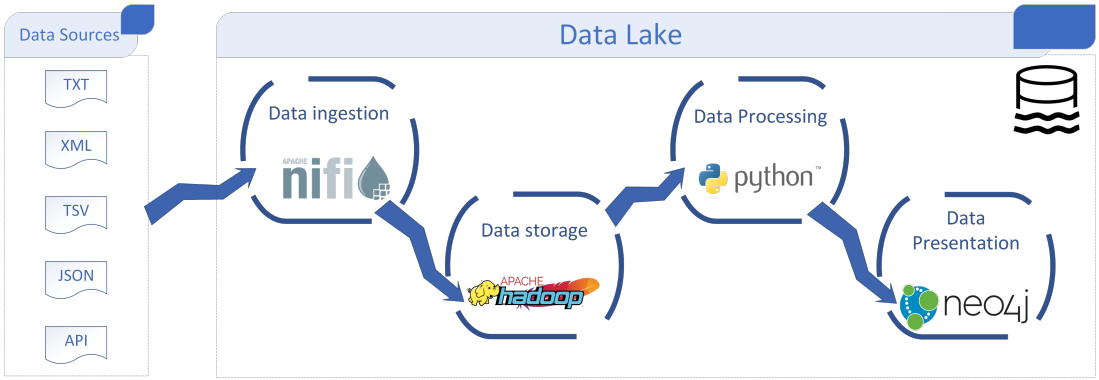

Fig. 1. Composition of the data lake

4.2 Data management

Data management includes data ingestion, storage, processing, and presentation, as illustrated in Figure 1. We ingested data into the data lake by creating processes in Apache Nifi. We create one process for each dataset, where the process searches for the dataset on HTTP (for MONDO, REACTOME, DISGENET, IID, and DRUGCENTRAL), or in a local folder (for OMIM and DRUGBANK, because they are not available directly due to the need of registering and licensing). Then we unzipped some of the datasets, and we renamed them all for standardization. Apache Nifi then sends the datasets to Hadoop, where they are stored in HDFS. For UNIPROT, we use the API directly on Jupyter. Then, we start data processing using the Python - Jupyter Notebook docker. Lastly, we create a graph visualization, based on Neo4J, to present the results.

4.3 Algorithm

We start the experiments by turning the datasets into Python Pandas dataframes. Pandas is a Python library for data analysis and manipulation. The standardization of the datasets as dataframes assure a unified entry for the algorithm, solving issues regarding one dataset being derived under one condition and the others being on other conditions. To create the DRUGBANK dataframe, we based on the solution provided by [37]. We also use other libraries, such as HDFS, that provide a pure HDFS client, bioservices that provide API access to UNIPROT, and the package *py_stringmatching* that implements the similarity metrics.

After creating the dataframes, one of the authors, a specialist in data science, analyzed the datasets to manually map the attributes candidates for points of integration. To do so, the specialist analyzed the names of the columns and a sample of data for each column, using data profiling techniques. The specialist took about four hours to finish this analysis, and we present the manual mapping in Figure 2. Figure 2 presents the manual data integration points, based on a graph visualization, where the nodes or vertices are the names of the dataframes, and the edges are the attributes' names. The orientation of the arrow indicates that, for instance, the attribute 'lbl' from the Mondo dataframe is a point of integration to the dataframe Disgenet, meaning that a percentage of 'lbl' is also present in another attribute of Disgenet. We developed this Figure to be later compared with the results of the algorithm we developed for data integration so that we could compare a user specialist analysis with the algorithm output.

Our algorithm is based on the concept of intersection or overlap between two attributes in sets of data. We first identify the unique values of each attribute for each dataset. Then

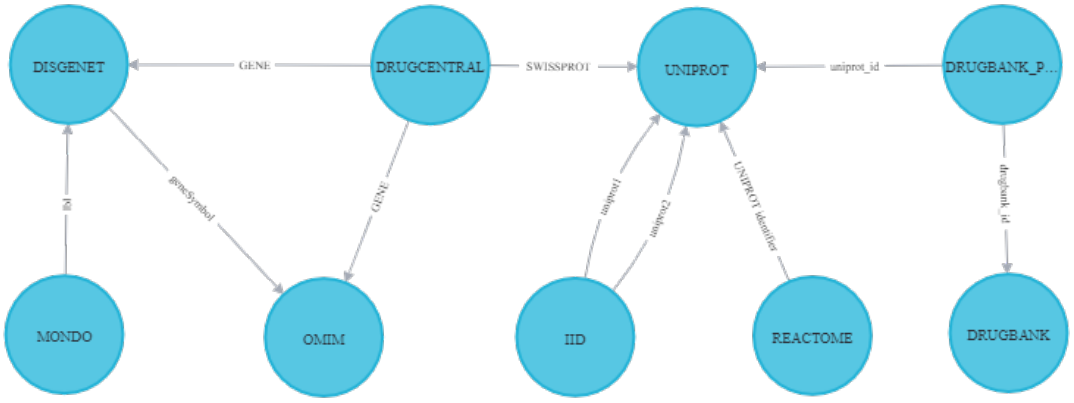

Fig. 2. Manually mapped integration

we compare each dataset attribute with all the other datasets' attributes to check if the unique values of the content of each attribute are contained in any other attributes of all of the other datasets. The attribute with fewer unique values indicates the orientation of the data integration. For instance, let us analyze the following case that includes dataframes (df) and attributes (att):

- df1['att01'] has 10 unique values;
- df2['att06'] has 20 unique values;
- 10 values from df1['att01'] are also present on df2['att06'].

In that case, we can notice that 100% of df1['att01'] are also present in df2['att06'], being that a good point for data integration. The notation would be: df1['att01'] $\longrightarrow$ df2['att06']. Regarding the minimum value for data intersection, we defined a threshold of 0.4 (in a range from 0–1)] to identify good integration points, but it is configurable according to the user's needs. It means that if two columns in a dataframe have 40% or more of data in common, the two columns are candidates for data integration, and the dataframes where the columns come from are integrable.

To define the best threshold for our experiments, we tested different values and compared the results with the specialist's analysis. We started with 0.9, and after each execution, we compared our results with the specialist's manual integration. When we noticed that the selected threshold retrieved at least all of the integration points defined by the specialist, we stopped decreasing the threshold, determining the value of 0.4.

Figure 3 details the activities diagram for the Algotrithm 1 we developed. Figure 3 shows that we can configure restrictions to select the attributes to be analyzed, such as the minimum number of unique values that an attribute must have to enter in the comparisons, and if we want to perform comparisons with attributes that contain only numeric values. Other restrictions include: removing attributes with only nulls or NaN and removing attributes with binary values (0 or 1, T or F). The binary values would not present real candidate points for data integration among the datasets since they mostly represent True or False values. For instance, the IID dataset presents dozens of attributes that are named after diseases, where the value = 0 corresponds to False and values = 1 corresponds to True (e.g.: 'bone disease', 'overnutrition', 'asthma', 'lymphoma').

The algorithm starts by creating dataframes for all the datasets, then it selects the first source dataframe, which will be compared to the other dataframes. Then the attributes of the source dataframe are compared with the attributes of the first candidate dataframe

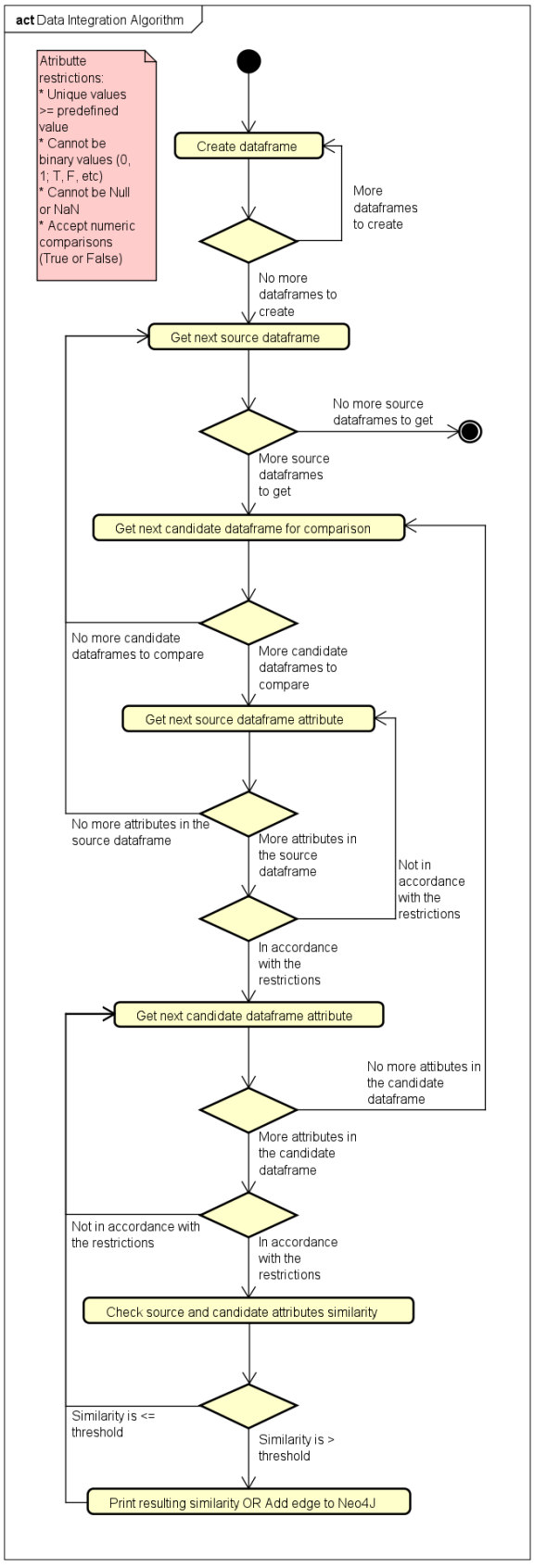

Fig. 3. Algorithm for data integration in UML Activity Notation

Algorithm 1 Pseudo-code for the data integration algorithm

Require: $_1$*datasets*, $_2$`minimum number of unique values`, $_3$`accept numeric comparisons` (*True or False*)
Ensure: `dataframeLeftName + columnName, dataframeRightName + columnName: Overlap:`*value* / `Jaccard:`*value* / `Sørensen-Dice:`*value* / `Tversky:`*value*

```
for dataframes do
    while columns in the dataframe = True do
        if compare numeric values = False then
            if value is numeric = True then
                next column in while
            end if
        end if
        if unique values ≤ predetermined value OR target column has already been compared = True then
            next column in while
        else
            if column values ≠ binary values then
                for dataframes + 1 do    ▷ repeats the same logic as the previous for for the next dataframe
                    ... for
                    if minimum number of unique values between compared columns ≠ 0 then
                        calculate Overlap, Jaccard, Sørensen-Dice, and Tversky
                        if Overlap, Jaccard, Sørensen-Dice, and Tversky > 0.4 then
                            return Output = Ensure
                        end if
                    end if
                end for
            end if
        end if
    end while
end for
```

to check the similarity. It happens until we do not have more source dataframes to be compared to the candidates.

Our algorithm also handles so that there are no redundant comparisons among dataframes and attributes. Firstly, we assure that a dataframe is not compared to itself by identifying its previously defined name in the algorithm. Secondly, when we compared each attribute of the first dataframe, we stored its description in a variable. Before comparing the dataframe's attribute with another, we check that there are no attributes with the same description. Therefore, we exclude the possibility of redundant comparisons between dataframes and attributes.

The algorithm returns a list having the names of the dataframes, attributes, and resulting values for the Szymkiewicz-Simpson overlap coefficient – Equation 2, which is the main result, compared to other similarity metrics (Jaccard – Equation 3, Sørensen-Dice – Equation 4, and Tversky – Equation 5). The resulting values for the similarity metrics range from 0 (attributes are not at all similar) to 1 (attributes contain the same data). Next, we present the equations related to the metrics, where X represents the attribute of the source dataframe and Y represents the attribute of the dataframe candidate to be compared.

The Overlap Equation calculates the size of the intersection divided by the smaller of the size of the two attributes or sets:

$$overlap(X, Y) = \frac{|X \bigcap Y|}{min(|X|, |Y|)} \tag{2}$$

The Jaccard measures the size of the intersection between two sets divided by the size of the union:

$$jaccard(X,Y) = \frac{|X \bigcap Y|}{|X \cup Y|} \tag{3}$$

The Sørensen-Dice similarity score returns twice the intersection divided by the sum of the cardinalities.

$$dice(X,Y) = \frac{2 \times |X \bigcap Y|}{|X| + |Y|} \tag{4}$$

The Tversky index is a generalization of the Sørensen-Dice's and the Tanimoto coefficient (aka Jaccard index) coefficient, but introduces the use of the parameters α and β, where $\alpha = \beta = 1$ produces the Tanimoto coefficient and $\alpha = \beta = 0.5$ produces the Sørensen–Dice coefficient:

$$tversky(X,Y) = \frac{|X \bigcap Y|}{|X \bigcap Y| + \alpha|X - Y| + \beta|Y - X|}; \alpha, \beta >= 0 \tag{5}$$

Our model also presents the option to insert nodes and edges in a Neo4J database to better visualize the relationships among the dataframes.

5 Results

After analyzing the first results presented by the algorithm, we identified that some suggested integration points are numeric values that, in our dataframes, do not represent actual data integration points. For instance:

- UNIPROT['Lenght'] it is the length of the canonical sequence and it varies from 3 to 4 numeric chars;
- OMIM['Entrez_Gene_ID'] the National Center for Biotechnology Information (NCBI) gene ID, values from 1 to 115029024;
- DRUGCENTRAL['STRUCT_ID'] the structure ID, and has values from 1 to 5390;
- DISGENET['YearInitial'] and DISGENET['YearFinal'] are years from 1924 to 2020;
- DISGENET['NofPmids'] the PubMed id, and has values from 1 to 67;
- DISGENET['NofSnps'] the Single nucleotide polymorphisms (SNP) id, has values from 1 to 284.

Because of that, we decided to add a parameter in the algorithm do set if we want to make numeric comparisons. We set the parameter to false since, in our case, it does not represent actual data integration points, but to be able to generalize for different domains and different types of datasets, that kind of comparison must be useful.

Regarding the similarity metrics, as Tversky, Sørensen-Dice, and Jaccard present correlated values for our data (Tverski's index with α and $\beta = 0.5$ was equal Sørensen-Dice and twice the Jaccard coefficient), we show only the Jaccard and the overlap values in Table 2.

After carefully comparing the Jaccard and Overlap results with the manual mapping and reviewing the actual dataframes, we identified that the Overlap provides better insights about the relationships that could be created among the dataframes.

For instance, for the relationship DISGENET["diseaseType"] and MONDO ["lbl"], the Jaccard index is equal to zero, while the Overlap is 0,667. We checked the data, and we really found a point for integration in that case. Another example is DRUGBANK["drugbank_id"] and DRUGBANK_PROTEIN["drugbank_id"], which represent the

Table 2. Final data integration mapping

DF1	Column DF1	DF2	Column DF2	Overlap	Jaccard
UNIPROT	Entry	← REACTOME	UNIPROT identifier	0,409	0,058
UNIPROT	Gene names	← OMIM	Approved_Symbol	0,466	0,015
UNIPROT	Gene names	← DISGENET	geneSymbol	0,483	0,009
UNIPROT	Gene names	← IID	symbol2	0,484	0,017
UNIPROT	Gene names	← DRUGCENTRAL	GENE	0,486	0,002
UNIPROT	Gene names	← IID	symbol1	0,488	0,017
DRUGCENTRAL	ACCESSION	→ DRUGBANK_PROTEIN	uniprot_id	0,488	0,018
UNIPROT	Organism	← DRUGBANK_PROTEIN	organism	0,499	0,206
DRUGCENTRAL	ACCESSION	→ IID	uniprot1	0,509	0,072
DRUGCENTRAL	ACCESSION	→ IID	uniprot2	0,510	0,071
DISGENET	geneSymbol	← DRUGCENTRAL	GENE	0,528	0,110
REACTOME	UNIPROT identifier	← DRUGBANK_PROTEIN	uniprot_id	0,546	0,030
REACTOME	UNIPROT identifier	← IID	uniprot2	0,565	0,107
REACTOME	UNIPROT identifier	← IID	uniprot1	0,567	0,105
DRUGBANK_PROTEIN	uniprot_id	→ IID	uniprot1	0,583	0,147
DRUGBANK_PROTEIN	uniprot_id	→ IID	uniprot2	0,590	0,147
OMIM	Approved_Symbol	← DRUGCENTRAL	GENE	0,609	0,082
DRUGCENTRAL	GENE	→ IID	symbol1	0,615	0,076
DRUGCENTRAL	GENE	→ IID	symbol2	0,617	0,075
REACTOME	UNIPROT identifier	← DRUGCENTRAL	ACCESSION	0,624	0,019
DISGENET	diseaseType	→ MONDO	lbl	0,667	0,000
REACTOME	Species	→ DRUGCENTRAL	ORGANISM	0,750	0,051
UNIPROT	Entry	← DRUGCENTRAL	ACCESSION	0,829	0,004
OMIM	Approved_Symbol	→ IID	symbol1	0,866	0,704
DISGENET	geneSymbol	→ IID	symbol1	0,880	0,464
OMIM	Approved_Symbol	→ IID	symbol2	0,885	0,717
DISGENET	geneSymbol	→ IID	symbol2	0,898	0,469
UNIPROT	Entry	← DRUGBANK_PROTEIN	uniprot_id	0,909	0,008
DISGENET	geneSymbol	→ OMIM	Approved_Symbol	0,915	0,536
UNIPROT	Entry	← IID	uniprot1	0,953	0,030
UNIPROT	Entry	← IID	uniprot2	0,960	0,031
DRUGBANK	drugbank_id	← DRUGBANK_PROTEIN	drugbank_id	1,000	0,579

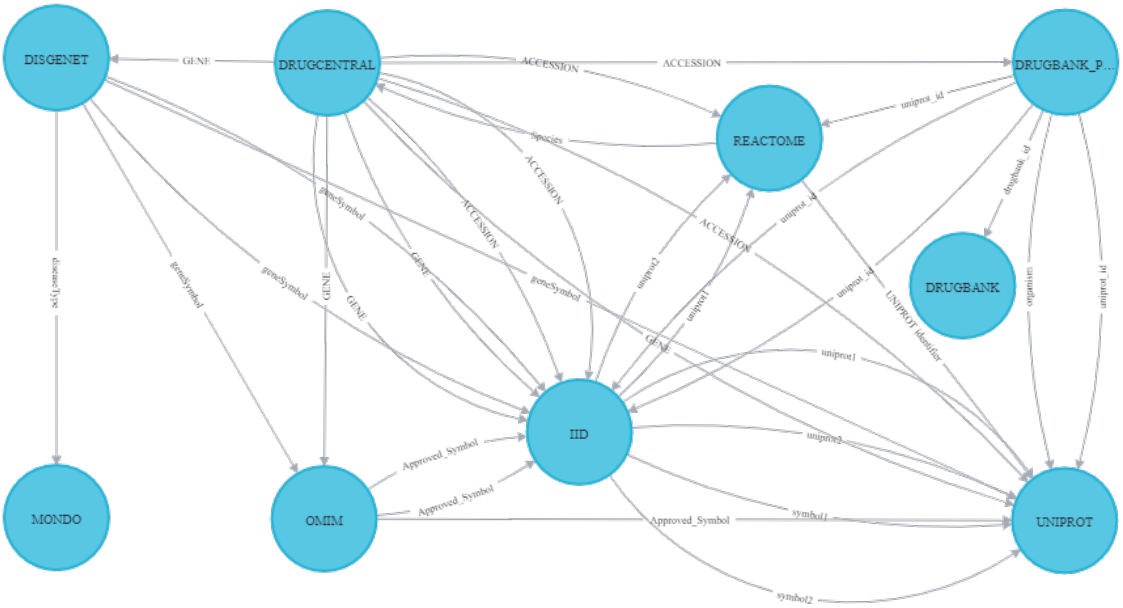

Fig. 4. Final data integration

same data according to the Overlap coefficient and to our manual analysis, and in this case, the Jaccard index is 0,579.

Hence, the Overlap coefficient seems to represent better how similar two attributes are and the level of data integration we could achieve if we integrate two dataframes using the top-ranked pairs of attributes indicated by the algorithm. Thus, we decided that in our case, it is better to use the Overlap Coefficient.

To better visualize the relationships between the dataframes, we create a database on Neo4J, where the nodes are the dataframes, and the edges are the name of the attributes. Figure 4 presents the final data integration resulting from our algorithm for the bioinformatics dataframes. In this graph visualization, similar to the manually mapped data integration visualization, the names of the dataframes are the vertices. The names of the attribute responsible for the integration are presented in the edges that connect the vertices. Figure 4 presents 9 nodes and 32 edges, some with only one edge between them (such as MONDO and DISGENET) and others having a high concentration of edges, such as IID, UNIPROT, and DRUGCENTRAL. The higher concentration of edges pointing to a dataframe means that the dataframe is referenced by a high number of other dataframes, meaning they represent an important point of integration.

When we manually mapped the points for integration, we identified 10 points (see Figure 2), while our model identifies 32 points, presented in Table 2. For instance, we manually mapped an integration between DRUGCENTRAL["SWISSPROT"] and UNIPROT ["ENTRY"], but our model shows that the coefficient for that integration is less than 0,4, while suggesting two better points: DRUGCENTRAL["GENE"] → UNIPROT[" Gene names "], with a overlap of 0,487, and DRUGCENTRAL["ACCESSION"] → UNIPROT["ENTRY"], with an overlap of 0,829.

Additionally, our model discovered 22 more paths of integration that were not manually identified, which we list above:

– From IID and: DISGENET, DRUGCENTRAL, OMIM, REACTOME, and DRUGBANK_PROTEIN

- From UNIPROT and: OMIM and DISGENET
- From REACTOME and: DRUGBANK_PROTEIN and DRUGCENTRAL
- From DRUGBANK_PROTEIN and DRUGCENTRAL

The scalability of the proposed solution takes place in terms of enabling comparisons between all attributes of all datasets. For example, the 19 attributes of DRUGCENTRAL are compared with the 253 attributes of the IID and so on, creating a bigger and bigger search space as we add more datasets for comparison.

Regarding the evaluation, we performed an analysis to answer the following question: *1) What is the average execution-time speedup provided by our model, including the data manipulation and algorithm?* We ran the model 10 times to get the average running time. It takes on average 2 hours and 30 minutes to run in the hardware we described in Section 4.1. Note that we run it in memory, in hardware with a humble configuration.

6 Discussion

We faced some challenges during the development and execution of our model. Initially, we had to elaborate on different ways of treating the datasets, as they had different data types. After this process, the researchers met to define the best way to carry out the comparison process. Effectively, the algorithm creation process started when we defined the four ways to calculate distances (Overlap, Jaccard, Sorensen, and Tversky's). Then, the initial algorithm implemented worked for most columns of the datasets. However, the algorithm generated errors, specifically for columns with information of the "JSON" or "XML" type, being corrected and treated soon afterward. After running the algorithm, we noticed that some of the comparisons generated 100% matches in many cases. Therefore, we verified that there were columns with information of binary values, which meant "False" or "True", but that was not necessarily relevant and similar to each other. We address this issue by removing columns with these data types from our comparison. We also skip Null and empty values, in the comparison steps. Furthermore, when we ran with all the datasets simultaneously, the initial version of the algorithm worked but took longer than we expected. Therefore, we performed refactoring in the algorithm, so we executed in the settings described in Section 4.1, we could obtain better results in a considerably shorter time.

The challenges we faced during algorithm development are all data-related. When we start data analysis with data pre-processing, the data must go through a cleaning phase, which could have ruled out some of the related challenges. However, one of the goals of the algorithm is to receive data from different formats with different types of attributes and be able to perform the necessary initial comparisons. In this way, we allow the algorithm to be executed even by people without specific knowledge in data processing, so they can and still obtain good results for their data integration.

Let us now discuss the utility of our model, by considering the data integration example in the field of bioinformatics. A data scientist has access to a data lake with the same bioinformatics datasets we worked on: OMIM, DISGENET, REACTOME, MONDO, DRUGBANK, IID, DRUGCENTRAL, and UNIPROT. The data scientist received the task to study neglected diseases, such as tuberculosis. To do so, it is necessary to explore the data related to the gene *inhA*, which is related to the organism *Mycobacterium tuberculosis*. Having those two pieces of information, it is easy to find the related data on UNIPROT. Actually, it may be on the top-5 results of a quick search on Google. But how will the person know if and how the data found in UNIPROT can be integrated with the other data sources, so they can find additional information? Well, usually the person

would have to put an effort into understanding the schema of all the datasets, analyze the data dictionary, a sample of data, and so on.

Using our data integration model, we will be able to see that UNIPROT is easily integrated with OMIM, DISGENET, IID, and DRUGCENTRAL by the *gene name*. By integrating with OMIM, we would have more details about genetic phenotypes related to the gene *inhA*; while DISGENET would bring the variants of the genes related to the tuberculosis disease. IID adds information about how a protein related to *inhA* (*Enoyl-[acyl-carrier-protein] reductase [NADH]*) interacts with other proteins. UNIPROT can also be integrated with REACTOME since REACTOME contains a field named *UNIPROT identifier*. Thus, we would have additional information about how the molecules interact in a cell to change the cell or create a certain product; for instance, turn genes on or off.

Additionally, integrating with DRUGCENTRAL would add information about interactions related to the drugs and tuberculosis. The integration with DRUGCENTRAL will allow integration with DRUGBANK, which brings supplementary information about the substance of the drugs and related products. For instance, we will find that *Pretomanid* is a medication for the treatment of tuberculosis. Finally, having the disease type from DISGENET, we could connect with the MONDO ontology, and learn about the different types of the disease, such as endocrine, esophageal, ocular, spinal tuberculosis, and others.

7 Conclusions

In this paper, we presented a model for automatized data integration in a Hadoop data lake, and we present experiments with eight well-known datasets from the bioinformatics domain, having different sizes and formats. We tested the similarity among the dataframes with different similarity measures, and we identified that The Overlap coefficient and Jaccard would be enough for us to validate our proposal.

Because the Overlap coefficient presented better results than the actual data and a specialists analysis, our experiments suggest that the Overlap coefficient is the best option for the in-memory set similarity approach we developed. Based on the Overlap coefficient, we found the top-k overlap set similarity that can help define data integration points for datasets in a data lake. For future work, we plan to implement text similarity strategies to magnify the reach of our results and increase the points for data integration based on semantic and syntactic.

Availability of data and materials

The data that support the findings of this study are available from:

- UNIPROT (UniProtKB - Reviewed (Swiss-Prot)). API available at [38].
- OMIM (genemap2). Available at [20], upon register and request. Release: File generated on 02/07/2020.
- DISGENET: Available at [39]. Release: version 7.0, January 2020.
- DRUGCENTRAL: Available at [40]. Release: 18/09/2020.
- IID: Available at [41]. Release: 2018-11.
- MONDO: Available at [42]. Release: v2021-01-15.
- REACTOME: Available at [43]. Release: Version 75, 07/12/2020.
- DRUGBANK: Available at [44], upon registration. Release: 5.1.8, 2021-01-03.

Restrictions apply to the availability of these data, which were used under license for the current study, and so are not all publicly available. Data are, however, available from the authors upon reasonable request and with permission of the owners, when necessary.

Acknowledgments

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES)* – Finance Code 001. We also thank Professor Anil Wipat and his team at Newcastle University for the support and advice in the early stages of the project and for providing us with the hardware for first creating the data lake.

References

1. D. B. Searls, "Data integration: challenges for drug discovery," *Nature reviews Drug discovery*, vol. 4, no. 1, pp. 45–58, 2005.
2. X. Lin, X. Li, and X. Lin, "A review on applications of computational methods in drug screening and design," *Molecules*, vol. 25, no. 6, p. 17, 2020.
3. A. Alserafi, A. Abelló, O. Romero, and T. Calders, "Towards information profiling: Data lake content metadata management," in *International Conference on Data Mining Workshops*, (Barcelona, ES), pp. 178–185, IEEE, 2016.
4. H. Dabbèchi, N. Z. Haddar, H. Elghazel, and K. Haddar, "Social media data integration: From data lake to nosql data warehouse," in *International Conference on Intelligent Systems Design and Applications*, (Online), pp. 701–710, 2020.
5. R. Hai, C. Quix, and C. Zhou, "Query rewriting for heterogeneous data lakes," in *European Conference on Advances in Databases and Information Systems*, (Budapest, HU), pp. 35–49, Springer, 2018.
6. Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada, "Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach," in *International Conference on Data Engineering*, (Chania, GR), pp. 456–467, IEEE, 2021.
7. P. Sawadogo and J. Darmont, "On data lake architectures and metadata management," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 97–120, 2021.
8. A. Y. Levy, *Logic-Based Techniques in Data Integration*, pp. 575–595. Boston, MA: Springer US, 2000.
9. M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, (New York, US), p. 233–246, Association for Computing Machinery, 2002.
10. E. Zhu, D. Deng, F. Nargesian, and R. J. Miller, "Josie: overlap set similarity search for finding joinable tables in data lakes," in *International Conference on Management of Data*, (Amsterdam, NL), pp. 847–864, ACM, 2019.
11. S. Ontañón, "An overview of distance and similarity functions for structured data," *Artificial Intelligence Review*, vol. 53, no. 7, pp. 5309–5351, 2020.
12. A. Tversky, "Features of similarity.," *Psychological review*, vol. 84, no. 4, p. 327, 1977.
13. T. J. Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. Copenhagen: I kommission hos Ejnar Munksgaard, 1948.
14. P. Jaccard, "Distribution comparée de la flore alpine dans quelques régions des alpes occidentales et orientales," *Bulletin de la Murithienne*, vol. XXXVII, pp. 81–92, 1902.
15. M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 19–28, 2016.
16. J. Couto, O. T. Borges, D. Ruiz, S. Marczak, and R. Prikladnicki, "A mapping study about data lakes: An improved definition and possible architectures," in *International Conference on Software Engineering and Knowledge Engineering*, (Lisbon, PT), pp. 453–458, KSI Research Inc., 2019.
17. R. R. Forster, *Hive on Spark and MapReduce: A methodology for parameter tuning*. Master thesis, NOVA Information Management School, Lisbon, PT, 2018.
18. Apache Software Foundation, "The Apache Software Foundation." `https://apache.org`, 2021. Accessed 25 Nov 2021.
19. A. Lesk, *Introduction to bioinformatics*. Oxford: Oxford university press, 2019.
20. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), "Online mendelian inheritance in man, omim®." `https://www.omim.org`, 2021. Accessed: 25 Nov 2021.
21. J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, pp. D845–D855, 2020.
22. B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 48, pp. D498–D503, 2020.

23. C. J. Mungall, J. A. McMurry, S. Köhler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad, *et al.*, "The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic acids research*, vol. 45, pp. D712–D722, 2017.
24. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, pp. D668–D672, 2006.
25. M. Kotlyar, C. Pastrello, Z. Malik, and I. Jurisica, "Iid 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species," *Nucleic acids research*, vol. 47, pp. D581–D589, 2019.
26. S. Avram, C. G. Bologa, J. Holmes, G. Bocci, T. B. Wilson, D.-T. Nguyen, R. Curpan, L. Halip, A. Bora, J. J. Yang, *et al.*, "Drugcentral 2021 supports drug discovery and repositioning," *Nucleic Acids Research*, vol. 49, pp. D1160–D1169, 2021.
27. U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, pp. D506–D515, 2019.
28. S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson, M. Young, and A. Wipat, "An integrated dataset for in silico drug discovery," *Journal of integrative bioinformatics*, vol. 7, pp. 15–27, 2010.
29. T. Sellis, D. Skoutas, and K. Staikos, "Database interoperability through web services and ontologies," in *International Conference on BioInformatics and BioEngineering*, (Athens, GR), pp. 1–5, IEEE, 2008.
30. J. Hendler, "Data integration for heterogenous datasets," *Big data*, vol. 2, pp. 205–215, 2014.
31. A. Petermann, M. Junghanns, R. Müller, and E. Rahm, "Graph-based data integration and business intelligence with biiig," *Proceedings of the VLDB Endowment*, vol. 7, pp. 1577–1580, 2014.
32. R. L. Bradshaw, R. Gouripeddi, and J. C. Facelli, "Concept bag: A new method for computing concept similarity in biomedical data," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, (Granada, ES), pp. 15–23, Springer, 2019.
33. Y. Zhang and Z. G. Ives, "Finding related tables in data lakes for interactive data science," in *International Conference on Management of Data*, (Portland, US), pp. 1951–1966, ACM, 2020.
34. H. Khalid and E. Zimányi, "Using rule and goal based agents to create metadata profiles," *Communications in Computer and Information Science*, vol. 1064, pp. 365–377, Sep, 2019.
35. D. Abadi, A. Ailamaki, D. Andersen, P. Bailis, M. Balazinska, P. Bernstein, ..., and D. Suciu, "The seattle report on database research," *SIGMOD Record*, vol. 48, p. 44–53, Dec, 2019.
36. H. Alrehamy and C. Walker, "SemLinker: automating big data integration for casual users," *Journal of Big Data*, vol. 5, pp. 1–14, Mar, 2018.
37. D. S. Himmelstein, "User-friendly extensions of the DrugBank database v1.0." `https://doi.org/10.5281/zenodo.45579`, Feb. 2016. Accessed 25 Nov 2021.
38. UniProt Consortium, "UniProt KB Reviewed (Swiss-Prot) dataset." `https://www.uniprot.org`, 2021. Accessed: 25 Nov 2021.
39. Integrative Biomedical Informatics Group, "DisGeNET curated gene-disease associations dataset." `https://www.disgenet.org/static/disgenet_ap1/files/downloads/curated_gene_disease_associations.tsv.gz`, 2021. Accessed: 25 Nov 2021.
40. S. Avram, C. G. Bologa, J. Holmes, G. Bocci, T. B. Wilson, D.-T. Nguyen, R. Curpan, L. Halip, A. Bora, J. J. Yang, *et al.*, "DrugCentral dataset." `https://drugcentral.org/download`, 2021. Accessed: 25 Nov 2021.
41. M. Kotlyar, C. Pastrello, Z. Malik, and I. Jurisica, "IID dataset." `http://iid.ophid.utoronto.ca/static/download/human_annotated_PPIs.txt.gz`, 2018. Accessed: 25 Nov 2021.
42. OBO Foundry, "Mondo dataset - json edition." `http://purl.obolibrary.org/obo/mondo/mondo-with-equivalents.json`, 2021. Accessed: 25 Nov 2021.
43. Reactome, "Reactome UniProt to pathways dataset." `https://reactome.org/download/current/UniProt2Reactome_All_Levels.txt`, 2021. Accessed: 25 Nov 2021.
44. OMx Personal Health Analytics, Inc., "DrugBank dataset." `https://go.drugbank.com/releases/latest`, 2021. Accessed 25 Nov 2021.

Authors

J Couto holds a degree in Information Systems (2012), a Master's in Computer Science (2018), and an MBA in Project Management (2016). She worked as a Project Manager, in distributed software projects in the health sector. Currently pursuing a Ph.D. in Computer Science at PUCRS, focusing on automating the integration of big data based on data profiling.

O. T. Borges is a Ph.D. student at PUCRS. He holds a degree in Information Systems (2015) and a Master's in Computer Science (2018). He is a member of the MuNDDoS Research Group (Distributed Software Development Research Group). His current research focus is supporting software development using Artificial Intelligence and Machine Learning techniques to Software Startups.

D. D. Ruiz holds a BS in Electrical Engineering from UFRGS (1983), a master's degree (1987), and a Ph.D. (1995) in Computer Science from the same university (UFRGS) and post-doctorate in Computer Science at Georgia Institute of Technology (2002). He is is a professor in the School of Technology at PUCRS, Brazil, where he leads the GPIN research group, acting primarily in the core of Machine Intelligence and Robotics. He has been working mainly in the areas of business process automation, non-conventional databases, bioinformatics, and database knowledge discovery (KDD).