Privacy-Preserving Big Hierarchical Data Analytics via Co-Occurrence Analysis

Alfredo Cuzzocrea[1,2][a] and Selim Soufargi[1][b]
[1] *iDEA Lab, University of Calabria, Rende, Italy*
[2] *Department of Computer Science, University of Paris City, Paris, France*

Keywords: Big Data, Privacy-Preserving Big Data, Big Hierarchical Data, Co-Occurrence Analysis, Multidimensional Big Data Analytics, Privacy-Preserving Multidimensional Big Data Analytics.

Abstract: Nowadays, *Big Data Analytics* is gaining the momentum in both the academic and industrial research communities. In this context, the issue of performing such a critical process under tight *privacy-preservation constraints* plays the critical role of "enabling technology". This paper, by perfectly aligning with the depicted paradigm, introduces and experimentally assesses *Drill*-CODA, an innovative framework that combines *drill-across multidimensional big data analytics and co-occurrence analysis to finally achieve privacy-preservation during the analytical phase.*

1 INTRODUCTION

Merging *privacy-preservation and big data analytics* (e.g., (Ram Mohan Rao et al., 2018; Tran and Hu, 2019)) is a first-quality research area that is gaining the attention from both the academic and industrial research communities. Indeed, while big data analytics (Russom, 2011; Tsai et al., 2015) offers noticeable tools for discovering hidden patterns and knowledge, severe *privacy breaches* are still possible, especially when related to personal information. *Aggregation* is a common practice to achieve privacy-preserving data analytics (e.g., (Singh and Kumar, 2023; Wei et al., 2024)) since aggregates remove details over personal data. This research line, in fact, has also originated a long series of research proposals in the context of *privacy-preserving OLAP* (e.g., (Agrawal et al., 2005)).

In the so-delineated research context, *big hierarchical data* (e.g., (Cuzzocrea et al., 2005; Ouazzani et al., 2021)) play a leading role, since they occur in a wide collection of application scenarios, ranging from censor data to logistic data, from geographic data to biological data, from sensor data to healthcare data, and so forth. It is worthy to consider that, in all these settings, big data analytics is a top-notch tool that is capable of enabling real actionable knowledge processing in the vest of a significant and valuable add-on for emerging applications.

This paper, by perfectly aligning with the depicted paradigm, introduces and experimentally assesses *Drill*-CODA, an innovative framework that combines *drill-across multidimensional big data analytics and co-occurrence analysis to finally achieve privacy-preservation during the analytical phase*. In *Drill*-CODA, the usage of co-occurrence analysis (e.g., (Honda et al., 2015; Wu et al., 2021)) combined with aggregates allows us to achieve an effective and powerful anonymization effect over big hierarchical data. The embedded drill-across query layer is used to magnify the capabilities of multidimensional big data analytics tools.

Figure 1 shows the *Drill*-CODA framework data processing workflow. It includes several layers/steps according to which input *raw data* are pre-processed at the *pre-processing layer*, even in order to discover the hidden hierarchies and to prepare them for the further *co-occurrence processing*. In the co-occurrence layer, co-occurrence analysis is performed, also to achieve the desired privacy-preserving effect (e.g., (Wang et al., 2018; Wang et al., 2020)). After this step, transformed co-occurrence data are aggregated according to their discovered hierarchies and a *multidimensional representation* is thus obtained. Suitable *integrated cubes* are consequently built and stored at this level. Finally, on top of the latter data cubes, a proper layer of *drill-across queries*

[a] https://orcid.org/0000-0002-7104-6415
[b] https://orcid.org/0009-0000-5476-9403

Cuzzocrea, A. and Soufargi, S.
Privacy-Preserving Big Hierarchical Data Analytics via Co-Occurrence Analysis.
DOI: 10.5220/0012767800003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 93-103
ISBN: 978-989-758-707-8; ISSN: 2184-285X

is executed in the vest of baseline tool for computing the final *privacy-preserving multidimensional big data analytics* (e.g., (Cuzzocrea, 2023)).

2 ANATOMY AND DATA PROCESSING STEPS OF *DRILL*-CODA

Here, we provide a description of the *Drill*-CODA steps: pre-processing, co-occurrence analysis, multi-dimensional aggregation, and drill-across querying.

In the *Drill*-CODA **pre-processing step**, the input hierarchical big datasets in $\mathcal{S}$ are treated for preparation for the next steps of the whole technique. First, we focus the attention on the anatomy of these datasets. Being hierarchical in nature, given a dataset $S_j \in \mathcal{S}$, some attributes $\mathcal{W}(S) = \{A_{k_0}, A_{k_1}, \ldots, A_{k_{|\mathcal{W}(S)|-1}}\} \in S_j$ play the role of *dimensions* while some other attributes $\mathcal{M}(S) = \{A_{h_0}, A_{h_1}, \ldots, A_{h_{|\mathcal{M}(S)|-1}}\} \in S_j$, such that $k_u \neq h_l \forall u \wedge l$, play the role of *measures* related to those dimensions. Given a dimension $A_{k_u} \in \mathcal{W}(S)$, a *dimensional hierarchy* $\mathcal{H}(A_{k_u})$ is defined on top of it, as follows: $\mathcal{H}(A_{k_u}) = \{l_{A_{k_u},0}, l_{A_{k_u},1}, \ldots, l_{A_{k_u},|\mathcal{H}(A_{k_u})|-1}\}$, such that $l_{A_{k_u},q}$ models a *hierarchical level* of $\mathcal{H}(A_{k_u})$, with $q \in \{0,1,\ldots,DEPTH(\mathcal{H}(A_{k_u}))-1\}$, where $DEPTH$ is a multidimensional operator that retrieves the depth of the hierarchy $\mathcal{H}(A_{k_u})$. However, as it will be clearer through the paper, while we keep in our model to respect the property of *autonomicity*, we do not process neither use the measures of datasets $S_j \in \mathcal{S}$ directly, since our framework is oriented to more advanced analytics.

In the pre-processing step, given a dataset $S_j \in \mathcal{S}$, we define: (*i*) a set of *target attributes* of interest for the analysis, namely $\mathcal{T}_{S_j} = \{T_{S_j,0}, T_{S_j,1}, \ldots, T_{S_j,|\mathcal{T}_{S_j}|-1}\}$, and the respective set of attribute values of interest for the analysis, namely $\mathcal{V}_{S_j} = \{V_{S_j,0}, V_{S_j,1}, \ldots, V_{S_j,|\mathcal{V}_{S_j}|-1}\}$, such $T_{S_j,k} = V_{S_j,k}, \forall k \in \{0,1,\ldots,|\mathcal{T}_{S_j}|-1 = |\mathcal{V}_{S_j}|-1\}$; (*ii*) a specific aggregate operator selected in the set $AO = \{SUM, COUNT, MIN, MAX, AVG\}$, which applies on top of the target attributes in $\mathcal{T}_{S_j}$; (*iii*) a set of *functional attributes* with respect to which the target attributes are analyzed, namely $\mathcal{F}_{S_j} = \{F_{S_j,0}, F_{S_j,1}, \ldots, F_{S_j,|\mathcal{F}_{S_j}|-1}\}$, such that $T_{S_j,k} \neq F_{S_j,h}, \forall k \neq h$.

Based on these definitions, we project S_j by target attributes in $\mathcal{T}_{S_j}$, and then we filter the obtained projected dataset by means of values in $\mathcal{V}_{S_j}$. After that, we apply the given aggregate operator in AO and we aggregate data of target attributes along *all* the hierarchies of dimensions in $\mathcal{W}(S_j)$. Of course, we aggregate the functional attributes in $\mathcal{F}_{S_j}$ as well. Formally, we denote the pre-processed dataset derived from S_j as S_j^{PP}, and we construct the set $\mathcal{S}^{PP} = \{S_0^{PP}, S_1^{PP}, \ldots, S_{|\mathcal{S}^{PP}|-1}^{PP}\}$.

In the *Drill*-CODA **co-occurrence analysis step**, the final goal is that of obtaining the privacy-preservation effect, since we apply a kind of *co-occurrence-based anonymization technique* that takes advantage from the multidimensional nature of target data. Before going into details, to become convinced about the approach, consider the following toy example. Let $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$ be two big healthcare datasets that store patient events about diseases, treatments, therapies and so forth, being the latter all *sensitive data* whose privacy should be preserved. Here, it is interesting and natural to analyze *correlations* that may exist among data $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$, in order, for instance, to discover *cross-therapies* performed by *different* hospitals over the *same* diseases, in order to ameliorate the effectiveness of combined therapies, perhaps obtained from the merging of therapies of different hospitals. In this case, let $Location$ and $Time$ be two *co-occurrence attributes*, both belonging to the schemes of $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$, respectively. Given a specific death event, for instance caused by cancer, it is possible to compute two different *co-occurrence datasets* from $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$, namely $CO[D_{i,\mathcal{H}}, D_{j,\mathcal{H}}, Location]$ and $CO[D_{i,\mathcal{H}}, D_{j,\mathcal{H}}, Time]$, respectively, such that $CO[D_{i,\mathcal{H}}, D_{j,\mathcal{H}}, Location]$ stores the death events of $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$ that refer to the *same* $Location$, while $CO[D_{i,\mathcal{H}}, D_{j,\mathcal{H}}, Time]$ stores the death events of $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$ that refer to the *same* $Time$, respectively. It should be noted that both the two co-occurrence attributes $Location$ and $Time$ model specific hierarchical levels of certain hierarchies associate to dimensions in both $D_{i,\mathcal{H}}$ and $D_{j,\mathcal{H}}$, respectively. Moreover, the co-occurrence analysis provides us with the desiderata privacy-preservation effect due to the fact that, when abstracted to the $Time$ level, e.g. $Year$, and the $Location$ level, e.g. $Country$, individual data are anonymized while aggregate data still suffice to the big data analytics purposes.

Formally, given the set of pre-processed hierarchical big datasets $\mathcal{S}^{PP} = \{S_0^{PP}, S_1^{PP}, \ldots, S_{|\mathcal{S}^{PP}|-1}^{PP}\}$ and a set of common co-occurrence attributes $\mathcal{A}_{\mathcal{S},CO} = \{A_{\mathcal{S},CO,0}, A_{\mathcal{S},CO,1}, \ldots, A_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}\} \in S_j \in \mathcal{S}$, such that $A_{\mathcal{S},CO,k} \in S_j^{PP}, \forall S_j^{PP} \in \mathcal{S}^{PP}, \forall k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\}$, we generate $|\mathcal{A}_{\mathcal{S},CO}|-1$ co-occurrence datasets, namely $CO_{\mathcal{S},CO} = \{C_{\mathcal{S},CO,0}, C_{\mathcal{S},CO,1}, \ldots, C_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}\}$,

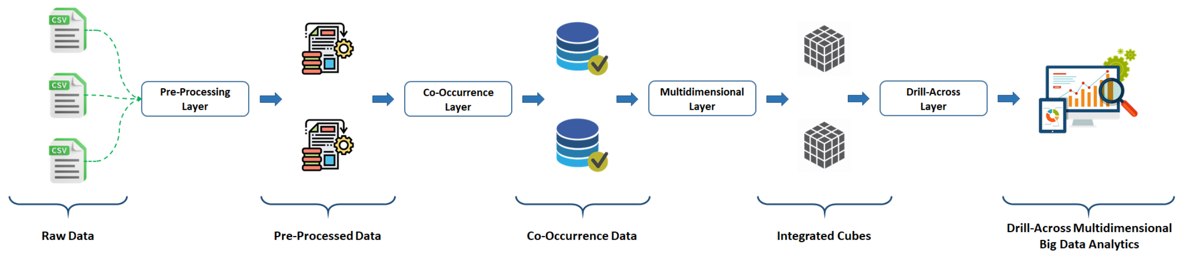

Figure 1: The *Drill*-CODA Framework Data Processing Workflow.

such that each dataset $C_{\mathcal{S},CO,k} \in CO_{\mathcal{S},CO}$ is defined as follows:

$$C_{\mathcal{S},CO,k} = \{A_{\mathcal{S},CO,k}, \langle F_{S_j,h}, \{AO_0(T_{S_j,0}), AO_1(T_{S_j,1}), \ldots, AO_{|\mathcal{T}_{S_j}|-1}(T_{S_j,|\mathcal{T}_{S_j}|-1})\}\rangle\}, \quad \forall k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\} \tag{1}$$

such that: (*i*) $A_{\mathcal{S},CO,k}$, where $k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\}$ denotes a co-occurrence attribute; (*ii*) $F_{S_j,h}$, where $h \in \{0,1,\ldots,|\mathcal{F}_{S_j}|-1\}$ denotes a functional attribute; (*iii*) AO_z, where $z \in \{0,1,\ldots,|AO|-1\}$, denotes an aggregate operator selected from the set AO.

To give an example, consider the schema of the first co-occurrence dataset, defined as follows: $\{Year, \langle Gender, COUNT(SkinCancer), COUNT(LungCancer), COUNT(DiabetesType1), COUNT(DiabetesType2)\rangle\}$. A possible instance is the following one: $\{2022, \{\langle F\text{-}Cancer, 35, 74\rangle, \langle M\text{-}Cancer, 37, 58\rangle, \langle M\text{-}Diabetes, 27, 51\rangle, \langle F\text{-}Diabetes, 43, 68\rangle\}\}$, which models the event that, during 2022, with *no* reference to the location, (*i*) a total of 109 female (F) patients died by cancer, specifically 35 of $SkinCancer$ and 74 of $LungCancer$; (*ii*) a total of 95 male (M) patients died by cancer, specifically 37 of $SkinCancer$ and 58 of $LungCancer$; (*iii*) a total of 78 male (M) patients died by diabetes, specifically 27 of $DiabetesType1$ and 51 of $DiabetesType2$; (*iv*) a total of 111 female (F) patients died by diabetes, specifically 43 of $DiabetesType1$ and 68 of $DiabetesType2$.

Similarly, consider the schema of the second co-occurrence dataset, defined as follows: $\{Country, \langle Gender, COUNT(SkinCancer), COUNT(LungCancer), COUNT(DiabetesType1), COUNT(DiabetesType2)\rangle\}$. A possible instance is the following one: $\{France, \{\langle M\text{-}Cancer, 28, 61\rangle, \langle F\text{-}Cancer, 35, 74\rangle, \langle M\text{-}Diabetes, 30, 63\rangle, \langle F\text{-}Diabetes, 43, 68\rangle\}\}$, which the event that, in *France*, with *no* reference to the time, (*i*) a total of 89 male (M) patients died by cancer, specifically 28 of $SkinCancer$ and 61 of $LungCancer$; (*ii*) a total of 109 female (F) patients died by cancer, specifically 35 of $SkinCancer$ and 74 of $LungCancer$; (*iii*) a total of 93 male (M) patients died by diabetes, specifically 30 of $DiabetesType1$ and 63 of $DiabetesType2$; (*iv*) a total of 111 female (F) patients died by diabetes, specifically 43 of $DiabetesType1$ and 68 of $DiabetesType2$.

From the examples above, it should be explicitly noted that, in our co-occurrence dataset, we group-by the aggregate values of the target attributes by means of the values of the functional attributes (e.g., *F-Cancer*: aggregate values of $COUNT(SkinCancer)$ and $COUNT(LungCancer)$ are grouped-by the gender of the patient F). This is due to the fundamental definition of co-occurrence analysis.

In the *Drill*-CODA **multidimensional aggregation step**, ad-hoc OLAP data cubes are built from the input co-occurrence datasets computed at the previous step (the co-occurrence analysis step). Given the input co-occurrence datasets $CO_{\mathcal{S},CO} = \{C_{\mathcal{S},CO,0}, C_{\mathcal{S},CO,1}, \ldots, C_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}\}$, we compute $|\mathcal{A}_{\mathcal{S},CO}|-1$ multidimensional OLAP data cubes as belonging to the set $\mathcal{DC}(CO_{\mathcal{S},CO}) = \{DC_{\mathcal{S},CO,0}, DC_{\mathcal{S},CO,1}, \ldots, DC_{\mathcal{S},CO,|\mathcal{DC}(CO_{\mathcal{S},CO})|-1}\}$, where $|\mathcal{A}_{\mathcal{S},CO}|-1 = |\mathcal{DC}(CO_{\mathcal{S},CO})|-1$, such that each data cube $DC_{\mathcal{S},CO,k} \in \mathcal{DC}(CO_{\mathcal{S},CO})$ is defined as follows:

$$DC_{\mathcal{S},CO,k} = \langle \{A_{\mathcal{S},CO,0}, A_{\mathcal{S},CO,1}, \ldots, A_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}\}, \{AO_0(T_{S_j,0}), AO_1(T_{S_j,1}), \ldots, AO_{|\mathcal{T}_{S_j}|-1}(T_{S_j,|\mathcal{T}_{S_j}|-1})\}\rangle, \quad \forall k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\} \tag{2}$$

such that: (*i*) $A_{\mathcal{S},CO,k}$, where $k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\}$ denotes a dimension (which corresponds to a co-occurrence attribute); (*ii*) AO_z, where $z \in \{0,1,\ldots,|AO|-1\}$, denotes an aggregate operator selected from the set AO; (*iii*) T_{S_k}, where $k \in \{0,1,\ldots,|\mathcal{T}_{S_j}|-1\}$, denotes a target attribute of interest for the analysis. It should be noted, here, that: (*i*) each OLAP data cube $DC_{\mathcal{S},CO,k} \in \mathcal{DC}(CO_{\mathcal{S},CO})$ is, formally, a *multiple-measure data cube*; (*ii*) the number of measures, which corresponds to the number of attributes of interest for the analysis, is the *same* for each OLAP data cube $DC_{\mathcal{S},CO,k} \in \mathcal{DC}(CO_{\mathcal{S},CO})$.

To give an example, consider a simple two-dimensional model. Here, let $\langle\{Year, Gender\text{-}$

Disease}, {$COUNT(\{SkinCancer, LungCancer\})$, $COUNT(\{DiabetesType1, DiabetesType2\})\}\rangle$ be the schema of the first (two-dimensional) OLAP data cube. A possible data cube cell instance is the following one: $\langle 2020, M\text{-}Cancer\rangle = \langle 32, 69\rangle$, which models the event that, during 2020, with *no* reference to the location, a total number of 32 male (*M*) patient died by *SkinCancer* and a total number of 69 male (*M*) patient died by *LungCancer*.

Similarly, let $\langle\{Country, Gender\text{-}Disease\}$, $\{COUNT(\{SkinCancer, LungCancer\}), COUNT(\{DiabetesType1, Diabetes\ Type2\})\}\rangle$ be the schema of the second (two-dimensional) OLAP data cube. A possible data cube cell instance is the following one: $\langle Italy, F\text{-}Diabetes\rangle = \langle 31, 55\rangle$, which models the event that, in *Italy*, with *no* reference to the time, a total number of 31 female (*F*) patient died by *DiabetesType*1 and a total number of 55 female (*F*) patient died by *DiabetesType*2.

In the *Drill*-CODA **drill-across querying step**, given the collection of OLAP data cubes $\mathcal{DC}(CO_{\mathcal{S},CO}) = \{DC_{\mathcal{S},CO,0}, DC_{\mathcal{S},CO,1}, \ldots, DC_{\mathcal{S},CO,|\mathcal{DC}(CO_{\mathcal{S},CO})|-1}\}$, computed at the previous step (the multidimensional aggregation step), we generate, for each data cube $DC_{\mathcal{S},CO,k} \in \mathcal{DC}(CO_{\mathcal{S},CO})$, a *full-dimensional drill-across query* $Q_{Q,CO,k}$, defined as follows:

$$\begin{aligned} Q_{\mathcal{S},CO,k} = \langle\{&[A_{\mathcal{S},CO,0}[0] : A_{\mathcal{S},CO,0}[|A_{\mathcal{S},CO,0}|-1]],\\ &[A_{\mathcal{S},CO,1}[0] : A_{\mathcal{S},CO,1}[|A_{\mathcal{S},CO,1}|-1]],\\ &\ldots,\\ &[A_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}[0] : A_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}\\ &[|A_{\mathcal{S},CO,|\mathcal{A}_{\mathcal{S},CO}|-1}|-1]]\}, AO_k(T_{S_j,k})\rangle\\ &\forall k \in \{0,1,\ldots,|\mathcal{DC}(CO_{\mathcal{S},CO})|-1\} \end{aligned} \tag{3}$$

such that: (*i*) $A_{\mathcal{S},CO,k}$, where $k \in \{0,1,\ldots,|\mathcal{A}_{\mathcal{S},CO}|-1\}$ denotes a dimension of $DC_{\mathcal{S},CO,k}$ (which corresponds to a co-occurrence attribute); (*ii*) $A_{\mathcal{S},CO,k}[0]$ denotes the *first* dimensional member in $A_{\mathcal{S},CO,k}$; (*iii*) $A_{\mathcal{S},CO,k}[|A_{\mathcal{S},CO,k}|-1]$ denotes the *last* dimensional member in $A_{\mathcal{S},CO,k}$; (*iv*) AO_z, where $z \in \{0,1,\ldots,|AO|-1\}$, denotes an aggregate operator selected from the set *AO*; (*v*) T_{S_k}, where $k \in \{0,1,\ldots,|\mathcal{T}_{S_j}|-1\}$, denotes a target attribute of interest for the analysis. It should be noted that the full-dimensional drill-across query $Q_{\mathcal{S},CO,k}$ spans *all* the dimensions of $DC_{\mathcal{S},CO,k}$ along *all* their dimensional domains.

By iterating the described procedure for each data cube $DC_{\mathcal{S},CO,k} \in \mathcal{DC}(CO_{\mathcal{S},CO})$, we obtain the so-called *full-dimensional drill-across query set* $\mathcal{Q}_{CO}(\mathcal{S}) = \{Q_{Q,CO,0}, Q_{Q,CO,1}, \ldots, Q_{Q,CO,|\mathcal{Q}_{CO}(\mathcal{S})|-1}\}$. After that, each drill-across query $Q_{Q,CO,k} \in \mathcal{Q}_{CO}(\mathcal{S})$ is executed against *all* the collection of OLAP data cubes $\mathcal{DC}(CO_{\mathcal{S},CO}) = \{DC_{\mathcal{S},CO,0}, DC_{\mathcal{S},CO,1}, \ldots, DC_{\mathcal{S},CO,|\mathcal{DC}(CO_{\mathcal{S},CO})|-1}\}$, thus finally originating the full-dimensional correlation set $\mathcal{D}_{CO}(\mathcal{S})$. From Section 1, remind that $\mathcal{D}_{CO}(\mathcal{S})$ stores collections of correlated aggregates.

To give an example, consider a simple two-dimensional model. Here, let $\langle\{Year, Gender\text{-}Disease\}, \{COUNT(\{SkinCancer, LungCancer\}), COUNT(\{DiabetesType1, DiabetesType2\})\}\rangle$ be the schema of the first (two-dimensional) OLAP data cube, and $\langle\{Country, Gender\text{-}Disease\}, \{COUNT(\{SkinCancer, LungCancer\}), COUNT(\{Diabetes\ Type1, Diabetes\ Type2\})\}\rangle$ be the schema of the second (two-dimensional) OLAP data cube, respectively. Let $\langle\{[2020 : 2023], [M\text{-}Cancer : F\text{-}Diabetes]\}, SUM\rangle$ be the input drill-across query against the two data cubes. The answer to the query is $\langle 358, 734\rangle$. The latter models the event that, from 2020 to 2023, a total number of 358 patients, with *no* reference to their sex, died by *Cancer* (including both *SkinCancer* and *LungCancer*), and a total number of 734 patients, with *no* reference to their sex, died by *Diabetes* (including both *DiabetesType*1 and *DiabetesType*2).

3 A COMPLETE *DRILL*-CODA CASE STUDY

In this Section, a complete example of *Drill*-CODA data processing workflow steps (see Section 1) is presented. For the sake of clarity and simplicity, we consider a simple but effective two-dimensional model. It is also worth noting that our approach is also valid for multidimensional models, as highlighted in Section 1. Specifically, our attention is directed toward the introduction of two synthetic hierarchical datasets, denoted as D_1 and D_2, designed to store disease-related information. Each record within these datasets represents a death event related to a particular disease. Figure 2 and Figure 3 show the structure and example record of D_1 and D_2, respectively.

For each dataset under consideration, we establish multidimensional hierarchies that provide a structured framework for organizing and analyzing the data. Specifically, both datasets feature two key hierarchies: a *temporal hierarchy* denoted as $\mathcal{H}(T) = Day \leftarrow Month \leftarrow Year$, capturing the temporal aspects of the data, and a *spatial hierarchy* denoted as $\mathcal{H}(S) = City \leftarrow Region \leftarrow Country$, representing the geographical dimensions. Beyond these fundamental hierarchies, additional attributes further enrich the datasets: (*i*) the attribute *Gender* serves to categorize

Attribute Name	Example Record
Day	15
Month	03
Year	2022
City	Nancy
Region	Grand-Est
Country	France
Gender	*F*
Disease	Cancer
Type	Lung

Figure 2: Structure and Example Record of the Dataset D_1 of the Case Study.

Attribute Name	Example Record
Day	18
Month	04
Year	2023
City	Florence
Region	Tuscany
Country	Italy
Gender	*F*
Disease	Diabetes
Type	Type 1

Figure 3: Structure and Example Record of the Dataset D_2 of the Case Study.

and model the gender of the patient; (*ii*) the attribute *Disease* encapsulates information about the disease affecting the patient; (*iii*) the attribute *Type* models the specific type of disease affecting the patient.

Indeed, the initial stage of *Drill*-CODA is devoted to pre-processing the input datasets, as described in Section 1. The functional property for D_1 and D_2 in our case study is *Gender*, whereas the target attribute is *Disease*. For our case study, we have used $COUNT$ as the aggregate operator. As a result, we utilize the values of *Cancer* for the attribute *Disease* and *Skin* and *Lung* for the (associated) attribute *Type* in D_1. Similarly, we use the values $Type\,1$ and $Type\,2$ of the (related) parameter *Type* and the value *Diabetes* of the attribute *Disease* to filter the data in D_2. In terms of the aggregate operator, we use $COUNT$ for the target attributes of both D_1 and D_2. Figure 4 shows the pre-processing for D_1 that generates the dataset $D_1[Cancer, \{Skin, Lung\}, COUNT]$ (here, *SC* denotes the attribute value *Skin* and *LC* denotes the attribute value *Lung*, respectively), while Figure 5 shows the pre-processing for D_2 that generates the dataset $D_2[Diabetes, \{Type\,1, Type\,2\}, COUNT]$ (here, $T1$ denotes the attribute value $Type\,1$ and $T2$ denotes the attribute value $Type\,2$, respectively).

Day	Month	Year	City	Region	Country	Gender	COUNT (SC)	COUNT (LC)
13	11	2020	Milan	Lombardy	Italy	*M*	32	69
10	05	2020	Munich	Bavaria	Germany	*F*	29	72
24	03	2021	Bordeaux	Nouvelle-Aquitaine	France	*M*	28	61
17	12	2021	Florence	Tuscany	Italy	*M*	12	44
15	02	2022	Nancy	Grand Est	France	*F*	35	74
09	09	2022	Dresden	Saxony	Germany	*M*	37	58

Figure 4: Dataset $D_1[Cancer, \{Skin, Lung\}, COUNT]$ after the Pre-Processing Step over D_1.

Day	Month	Year	City	Region	Country	Gender	COUNT (T1)	COUNT (T2)
13	11	2020	Rome	Lazio	Italy	*M*	29	61
10	05	2021	Leipzig	Saxony	Germany	*F*	25	68
24	03	2021	Lille	Haut-de-France	France	*M*	30	63
17	12	2022	Stuttgart	Baden-Wurttemberg	Germany	*M*	27	51
15	02	2022	Paris	Ile-de-France	France	*F*	43	68
09	09	2023	Naples	Campania	Italy	*F*	31	55

Figure 5: $D_2[Diabetes, \{Type\,1, Type\,2\}, COUNT]$ after the Pre-Processing Step over D_2.

The *Drill*-CODA approach requires the co-occurrence analysis to be conducted following the pre-processing stage (see Section 1). In Section 2, pre-processed datasets are used to find frequent co-occurrence attributes based on analytic goals, resulting in relevant co-occurrence datasets. Specifically, in this case study and for the purpose of ensuring high privacy-preservation, we select *Year* and *Country* as co-occurrence attributes, according to the guidelines discussed in Section 2. Figure 6 and Figure 7 show the co-occurrence dataset originated from the co-occurrence analysis on the (pre-processed) datasets $D_1[Cancer, \{Skin, Lung\}, COUNT]$ and $D_2[Diabetes, \{1, 2\}, COUNT]$ over *Year*, and the (pre-processed) datasets $D_1[Cancer, \{Skin, Lung\}, COUNT]$ and $D_2[Diabetes, \{1, 2\}, COUNT]$ over *Country*, respectively.

Year	Co-Occurrence Data
2020	$\{\langle M - Cancer, 32{,}69\rangle, \langle F - Cancer, 29{,}72\rangle, \langle M - Diabetes, 29{,}61\rangle\}$
2021	$\{\langle M - Cancer, 40{,}105\rangle, \langle F - Diabetes, 25{,}68\rangle, \langle M - Diabetes, 30{,}63\rangle\}$
2022	$\{\langle F - Cancer, 35{,}74\rangle, \langle M - Cancer, 37{,}58\rangle, \langle M - Diabetes, 27{,}51\rangle, \langle F - Diabetes, 43{,}68\rangle\}$
2023	$\{\langle F - Diabetes, 31{,}55\rangle\}$

Figure 6: Co-Occurrence Dataset Generated from Datasets $D_1[Cancer, \{Skin, Lung\}, COUNT]$ and $D_2[Diabetes, \{1, 2\}, COUNT]$ over *Year*.

Country	Co-Occurrence Data
Italy	$\{\langle M - Cancer, 44{,}113\rangle, \langle M - Diabetes, 29{,}61\rangle, \langle F - Diabetes, 31{,}55\rangle\}$
Germany	$\{\langle F - Cancer, 29{,}72\rangle, \langle M - Cancer, 37{,}58\rangle, \langle F - Diabetes, 25{,}68\rangle, \langle M - Diabetes, 27{,}51\rangle\}$
France	$\{\langle M - Cancer, 28{,}61\rangle, \langle F - Cancer, 35{,}74\rangle, \langle M - Diabetes, 30{,}63\rangle, \langle F - Diabetes, 43{,}68\rangle\}$

Figure 7: Co-Occurrence Dataset Generated from Datasets $D_1[Cancer, \{Skin, Lung\}, COUNT]$ and $D_2[Diabetes, \{1, 2\}, COUNT]$ over *Country*.

Figure 8 presents the *Time* co-occurrence analytics over the co-occurrence dataset shown in Figure 6, while Figure 9 presents the *Location* co-occurrence analytics over the co-occurrence dataset shown in Figure 7, respectively.

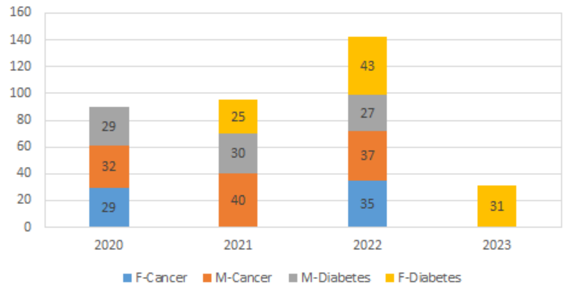

Figure 8: *Time* Co-Occurrence Analytics over Co-Occurrence Dataset of Figure 6.

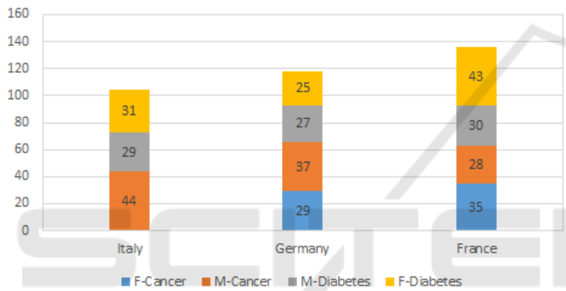

Figure 9: *Location* Co-Occurrence Analytics over Co-Occurrence Dataset of Figure 7.

Figure 8 and Figure 9 show that the count of deaths per gender and per disease on the Y axis and either the year or the location, respectively, on X axis. Detailed count per month (see Figure 8) or per city (see Figure 9) is therefore not displayed, and the data are anonymized up to the highest hierarchical level of the *time*/*location* attributes. The highest location co-occurrences has happened in France, with more than 130 cases across the four possible values of $Gender - Disease$ attribute, and the least were in Italy, with roughly a bit more than 100 death cases and where no female has died of cancer. Whereas for the *time* co-occurrences, the highest count of death cases is registered for the year 2022 and the least count is registered for the year 2023, where only female death cases from diabetes were registered.

Following the acquisition of co-occurrence data, the subsequent step involves computing suitable OLAP data cubes for supporting big data analytics (see Section 1). In our specific case study, utilizing the two co-occurrence datasets generated during the preceding stage of *Drill*-CODA, we proceed with the creation of two-dimensional OLAP data cubes. The initial cube, denoted as A_1, is defined as $A_1 = \langle\{Year, Gender\text{-}Disease\}, \{COUNT(\{Skin, Lung\}), COUNT(\{Type1, Type2\})\}\rangle$ (see Figure 10). This data cube encapsulates the temporal dimension ($Year$) and the composite $Gender - Disease$ category. Simultaneously, the second OLAP data cube A_2 is defined as $A_2 = \langle\{Country, Gender\text{-}Disease\}, \{COUNT(\{Skin, Lung\}), COUNT(\{Type1, Type2\})\}\rangle$ (see Figure 11), which delves into the geographical aspect by incorporating the $Country$ dimension alongside the $Gender - Disease$ attribute.

Gender / Year	$M-Cancer$	$F-Cancer$	$M-Diabetes$	$F-Diabetes$
2020	(32,69)	(29,72)	(29,61)	
2021	(40,105)		(30,63)	(25,68)
2022	(37,58)	(35,74)	(27,51)	(43,58)
2023				(31,55)

Figure 10: Two-Dimensional OLAP Data Cube $A_1 = \langle\{Year, Gender\text{-}Disease\}, \{COUNT(\{Skin, Lung\}), COUNT(\{Type1, Type2\})\}\rangle$.

Gender / Country	$M-Cancer$	$F-Cancer$	$M-Diabetes$	$F-Diabetes$
Italy	(44,113)		(29,61)	(31,55)
Germany	(37,58)	(29,72)	(27,51)	(25,68)
France	(28,61)	(35,74)	(30,63)	(43,68)

Figure 11: Two-Dimensional OLAP Data Cube $A_2 = \langle\{Country, Gender\text{-}Disease\}, \{COUNT(\{Skin, Lung\}), COUNT(\{Type1, Type2\})\}\rangle$.

As shown in Figure 10 and Figure 11, we can notice that the dimensions of the OLAP data cubes are ordered according to a certain *topological ordering*. This conclusion is influenced by considering the data organization and OLAP query performance.

Figure 12 shows the *Time two-dimensional* co-occurrence analytics derived from the OLAP data cube in Figure 10, while Figure 13 shows the *Location two-dimensional* co-occurrence analytics derived from the OLAP data cube in Figure 11, respectively. Here, for each time/location index (e.g., 2020 or Germany), we show both values of the couple of measures representing the count of deaths by the sub-type of the diseases.

The final goal of our *Drill*-CODA framework consists of performing and building the full-dimensional correlation set $\mathcal{D}_{CO}(\mathcal{S})$ (see Section 1). This latter is tailored to store sets of correlated aggregates retrieved from the execution of a suitable set of drill-across queries along *all* the hierarchical dimensions defined on the input set of hierarchical big datasets $\mathcal{S}$,

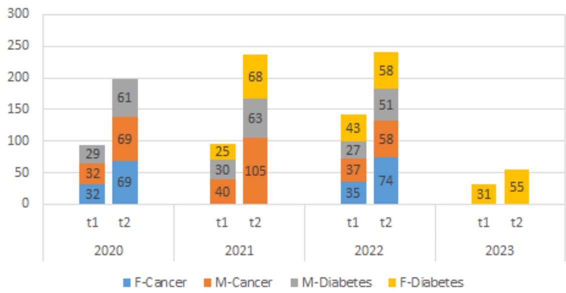

Figure 12: *Time* Two-Dimensional Co-Occurrence Analytics derived from the OLAP Data Cube in Figure 10.

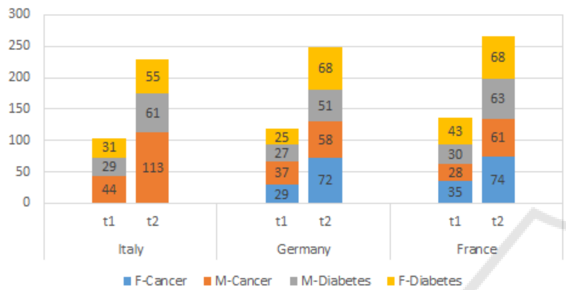

Figure 13: *Location* Two-Dimensional Co-Occurrence Analytics derived from the OLAP Data Cube in Figure 11.

taking as input the ad-hoc OLAP data cubes built at the third step of the *Drill*-CODA's methodology.

The full-dimensional correlation set $\mathcal{D}_{CO}(\mathcal{S})$ is computed by executing *all* the sets of admissible *full-dimensional* drill-across queries over datasets in $\mathcal{S}$, along *all* their dimensional domains (see Section 2). Figure 14 shows the full-dimensional correlation set $\mathcal{D}_{CO}(\{D_1, D_2\})$ for the running case study.

$\mathcal{D}_{CO}(\mathcal{S})$, being $\mathcal{S} = \{D_1, D_2\}$, according to what described in Section 2, is computed by executing *all* the set of admissible *full-dimensional* drill-across queries over datasets in $\mathcal{S}$, along *all* their dimensional domains. Figure 14 shows the full-dimensional correlation set $\mathcal{D}_{CO}(\{D_1, D_2\})$ for the running case study.

Country	Year	$M-Cancer$	$F-Cancer$	$M-Diabetes$	$F-Diabetes$
Italy	2020	⟨76,182⟩	⟨29,72⟩	⟨58,122⟩	⟨31,55⟩
	2021	⟨84,218⟩		⟨59,124⟩	⟨56,123⟩
	2022	⟨81,171⟩	⟨35,74⟩	⟨56,112⟩	⟨74,113⟩
	2023	⟨44,113⟩		⟨29,61⟩	⟨62,110⟩
Germany	2020	⟨69,127⟩	⟨58,144⟩	⟨56,112⟩	⟨25,68⟩
	2021	⟨77,163⟩	⟨29,72⟩	⟨57,114⟩	⟨50,136⟩
	2022	⟨74,116⟩	⟨64,146⟩	⟨54,102⟩	⟨68,126⟩
	2023	⟨37,58⟩	⟨29,72⟩	⟨27,51⟩	⟨56,123⟩
France	2020	⟨60,130⟩	⟨64,146⟩	⟨59,124⟩	⟨43,68⟩
	2021	⟨68,166⟩	⟨35,74⟩	⟨60,126⟩	⟨68,136⟩
	2022	⟨65,119⟩	⟨70,148⟩	⟨57,114⟩	⟨86,126⟩
	2023	⟨28,61⟩	⟨35,74⟩	⟨30,63⟩	⟨74,123⟩

Figure 14: Full-Dimensional Correlation Set $\mathcal{D}_{CO}(\{D_1, D_2\})$ for the Running Case Study.

In this research, we conduct a correlation analysis over the full-dimensional correlation set $\mathcal{D}_{CO}(\{D_1, D_2\})$ via two widely used correlation metrics (i.e., Pearson correlation coefficient and the Spearman correlation coefficient) (Corder and Foreman, 2014).

Furthermore, for *each* correlated aggregate pair $\langle M_1, M_2 \rangle$ of the full-dimensional correlation set $\mathcal{D}_{CO}(\{D_1, D_2\})$, we compute the Pearson correlation coefficient and the Spearman correlation coefficient in order to obtain the so-called *full-dimensional Pearson correlation set*, denoted by $\mathcal{P}_{CO}(\{D_1, D_2\})$, and the so-called *full-dimensional Spearman correlation set*, denoted by $\mathcal{S}_{CO}(\{D_1, D_2\})$, respectively.

Indeed, Figure 15 and Figure 16 show the full-dimensional Pearson correlation set $\mathcal{P}_{CO}(\{D_1, D_2\})$ and the full-dimensional Spearman correlation set $\mathcal{S}_{CO}(\{D_1, D_2\})$ for the running case study, respectively.

Year / Country	2020	2021	2022	2023
Italy	1	1	0.9	1
Germany	0.9	0.9	0.3	0.9
France	1	0.9	0.3	1

Figure 15: Full-Dimensional Pearson Correlation Set $\mathcal{P}_{CO}(\{D_1, D_2\})$ for the Running Case Study.

Year / Country	2020	2021	2022	2023
Italy	0.8	1	1	0.8
Germany	0.8	0.8	0.2	0.8
France	1	1	0.8	1

Figure 16: Full-Dimensional Spearman Correlation Set $\mathcal{S}_{CO}(\{D_1, D_2\})$ for the Running Case Study.

4 *DRILL*-CODA CLOUD-BASED REFERENCE ARCHITECTURE

In this Section, we introduce the Cloud-based reference architecture for the proposed *Drill*-CODA framework. We start by elucidating the underlying motivation for a real-world case study of our technique and highlighting how *Drill*-CODA can be successfully used in the context of big data analytics platforms.

Modern big data analytics applications usually run on top of massive, large-scale big data repositories. As a consequence, there is a need for accessing, processing, and analyzing such repositories via both well-consolidated big data management and analytics techniques and well-established Cloud-based big data processing platforms, such as *Hadoop*, *Spark*, and *Kylin*.

In reply to these clear requirements, *Drill*-CODA must be deployed in a naive big data environment, as to take advantage of high-computation capabilities, scalability, virtualization, parallel/distributed executions, in-memory partial computations, and so forth. This evidence is stirred-up by the fact that *Drill*-CODA mostly processes multidimensional big data, hence, it can easily incur in the so-called *curse of dimensionality* problem (e.g., (Cuzzocrea et al., 2003)), meaning that performance of algorithms over multidimensional data decreases when the number of dimensions of input datasets increases. As a consequence, our study explores the anatomy and the functionalities of the big-data-aware *Drill*-CODA deployment. Figure 17 shows the Cloud-based *Drill*-CODA reference architecture.

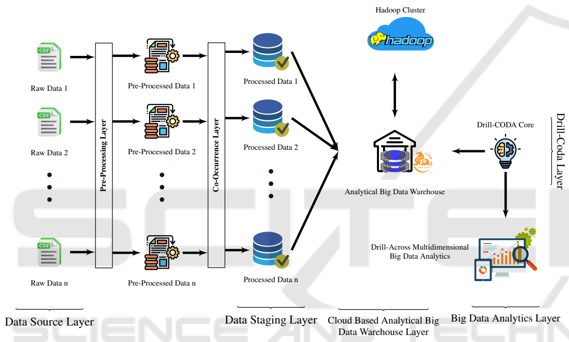

Figure 17: The Cloud-Based *Drill*-CODA Reference Architecture.

As shown in Figure 17, the Cloud-based *Drill*-CODA reference architecture includes the following layers:

1. ***Data Source Layer:*** In this layer, the original data sources of our Cloud-based *Drill*-CODA framework are fed as input to our enabling tool. Data, as collected from their sources (web, repositories, and so forth), are used as main entry for our data flow. Depending on their format and structure, which should be "unified" for subsequent processing, we apply cleansing and formatting transformations on them before considering them ready for the next data staging phase.
2. ***Pre-Processing Layer.*** Here, normalized data sources are pre-processed according to the *Drill*-CODA paradigm (see Section 2). This calls for a pre-processing step to cleanse and reformat data columns when needed, and above all, the crafting of data for the respective co-occurrence attributes, so that a valid drill-down operation could later be applied to the OLAP cubes to analyze. Also, aggregation along hierarchies is performed.
3. ***Co-Occurrence Layer.*** Here, the *Co-Occurrence Layer* supports our co-occurrence analysis (see Section 2). Our main goal through this phase is to ensure that co-occurrence attributes are present and allow the creation of a consequent hierarchy later-on for our multidimensional analysis. The co-occurrence aggregate data are provided as final output.
4. ***Data Staging Layer.*** In this layer, we materialize the co-occurrence data into suitable data structures, on top of which multidimensional analysis is later performed. This step is required to prepare the data for querying in highly-multidimensional fashion and make the data (type and format essentially) suitable for deployment onto the *data warehouse solutions*.
5. ***Cloud-Based Analytical Big Data Warehouse Layer.*** In this layer, thanks to the *Kylin* OLAP framework and its interoperability with *Hadoop*, multidimensional data are aggregated on top of staging co-occurrence data in a *MapReduce* fashion. Indeed, Kylin is a big data platform for data warehousing and OLAP that integrates a Spark-based OLAP engine needed for the Hadoop MapReduce parallel data processing. In fact, Kylin is capable of integrating, deploying, and processing a high number of cubes in a concurrent manner through Hadoop. In our case study, we use Kylin MDX to query the cube using *Multidimensional Expressions* (MDX). Indeed, after including the staged data sources and after creating the data model of the cube as well as the deployment of the cube in Kylin, the tool enables the querying through MDX using a third-party *Business Intelligence* tool such as Tableau or Excel. Figure 18 and Figure 19 show the deployment of cubes in Kylin and Kylin MDX, respectively.

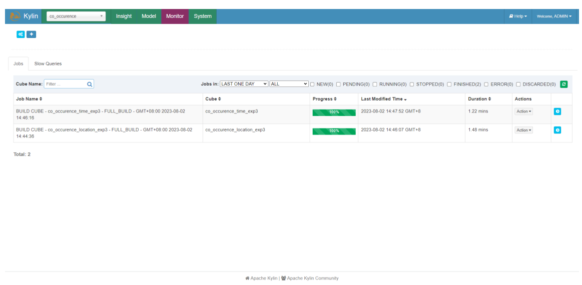

Figure 18: Deployed Cubes in Kylin.

An example of MDX query, we are using the extract the data from one cube is shown in Figure 20.

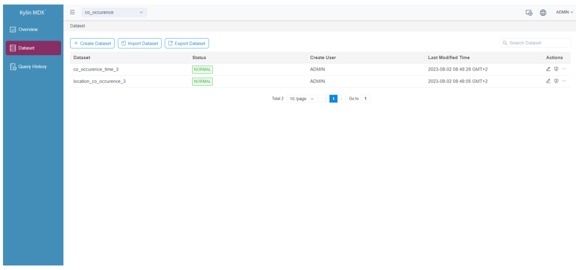

Figure 19: Deployed Cubes in Kylin MDX.

```
WITH
MEMBER  MEASURES.COUNT1 AS [Measures].[Count1]
MEMBER  MEASURES.COUNT2 AS [Measures].[Count2]
SELECT { MEASURES.COUNT1, MEASURES.COUNT2 }
     ON COLUMNS,
NON EMPTY{(
DRILLDOWNLEVEL({ [Location_Co_Occurence]
.[Hierarchy].[City_Name] }
     ),[Location_Co_Occurence]
.[Substance_Gender].[Substance_Gender])}
ON ROWS
FROM [location_co_occurence_cube]
```

Figure 20: MDX Query to Drill-Down from *Region* → *Country* → *City*.

6. ***Drill-CODA Layer.*** In the *Drill-CODA Layer*, the core components of *Drill*-CODA run in order to derive drill-across multidimensional big data analytics over big co-occurrence aggregate hierarchical data, according to the main guidelines proposed by our research (see Section 2).
7. ***Big Data Analytics Layer.*** Here, the final desiderata big data analytics applies, in order to provide useful and actionable knowledge from large-scale big data repositories, mostly by focusing the attention on the full-dimensional correlation pattern discovery (see Section 3).

5 EXPERIMENTAL ANALYSIS AND RESULTS

In this Section, we present our experimental assessment of the proposed *Drill*-CODA framework. This involves conducting several experimental tests over large-scale real-life datasets in order to evaluate the performance and capabilities of the framework.

As regards datasets, we deliberately selected different real-life datasets, as to give more reliability to the scope and effectiveness of our experimental campaign. In compliance with the primary objectives of the framework (see Section 2), we perform our evaluation based on co-occurrence analysis.

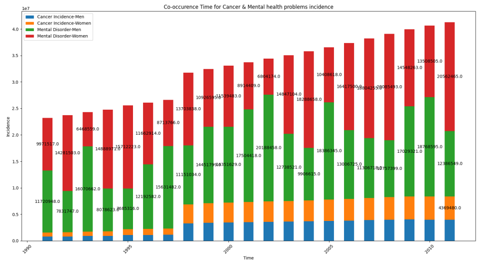

Figure 21: *Time* Co-Occurrence Analysis over the Cancer-Incidence/Mental-Disorders Experimental Setup.

In more details, we focus on the *correlation between cancer incidence and mental disorders*. Here, we used the following real-life datasets: (*i*) **Cancer Incidence (CI5Plus)**: the *CI5Plus* database contains updated annual incidence rates for 124 selected populations from 108 cancer registries published in *CI5Plus*, for the longest period available (up to 2012), for all cancers and 28 major types (Organization, 2023); (*ii*) **Mental Disorders**: this dataset contains informative data from Countries across the globe about the prevalence of mental health disorders, including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression and alcohol use disorders (Devastator, 2023).

In our evaluation, we conduct a *co-occurrence analysis* (i.e., time and location co-occurrence) over the previously described experiment. Here, we display the findings of our investigation that were generated using *Python/Matplotlib* library. Therefore, let us notice that co-occurrence data is plotted in an anonymized manner, since only the *Year* (*Region*, respectively) attribute numbers are depicted, being those attributes the higher level of the time and the location hierarchies.

For the time co-occurrence analysis (see Figure 21), a spike in cancer incidence is noticeable starting from year 1998, while mental disorders counting was highly fluctuating for both men and women. On the other hand, Figure 22 shows the location co-occurrence analysis over our experimental setup. It should be noted that a higher number of cancer and mental disorders were still registered in Asia & Pacific and Europe regions, while Africa had low numbers of incidence of the considered health diseases.

6 CONCLUSIONS AND FUTURE WORK

This paper has presented and experimentally assessed *Drill*-CODA, a framework designed for supporting drill-across multidimensional big data analytics on large-scale co-occurrence aggregate hierarchical data.

Future work is mainly oriented towards extending our proposed framework by means of innovative characteristics of the emerging big data processing paradigm, such as: (*i*) *management of uncertain and imprecise hierarchical data* (e.g., (Burdick et al., 2007)); (*ii*) *anomaly detection* (e.g., (Langone et al., 2020)); (*iii*) *inference detection* (e.g., (Chow et al., 2008)); (*iv*) *explainability* (e.g., (Aghaeipoor et al., 2022)); (*v*) *visualization* (e.g., (Cuzzocrea and Mansmann, 2009; Barkwell et al., 2018)).

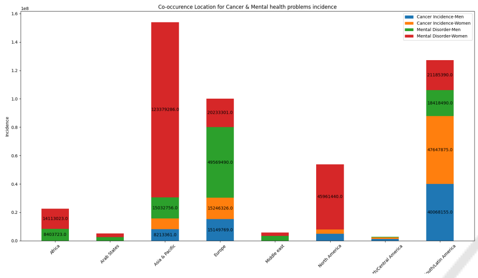

Figure 22: *Location* Co-Occurrence Analysis over the Cancer-Incidence/Mental-Disorders Experimental Setup.

ACKNOWLEDGEMENTS

This work was funded by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5 (Directorial Decree n. 2021/3277) - project Tech4You n. ECS0000009.

REFERENCES

Aghaeipoor, F., Javidi, M. M., and Fernández, A. (2022). IFC-BD: an interpretable fuzzy classifier for boosting explainable artificial intelligence in big data. *IEEE Trans. Fuzzy Syst.*, 30(3):830–840.

Agrawal, R., Srikant, R., and Thomas, D. (2005). Privacy preserving OLAP. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, pages 251–262. ACM.

Barkwell, K. E., Cuzzocrea, A., Leung, C. K., Ocran, A. A., Sanderson, J. M., Stewart, J. A., and Wodi, B. H. (2018). Big data visualisation and visual analytics for music data mining. In *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*, pages 235–240. IEEE Computer Society.

Burdick, D., Deshpande, P. M., Jayram, T. S., and Al., E. (2007). OLAP over uncertain and imprecise data. *VLDB J.*, 16(1):123–144.

Chow, R., Golle, P., and Staddon, J. (2008). Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 893–901.

Corder, G. W. and Foreman, D. I. (2014). *Nonparametric Statistics: A Step-by-Step Approach*. Wiley.

Cuzzocrea, A. (2023). A reference architecture for supporting multidimensional big data analytics over big web knowledge bases: Definitions, implementation, case studies. *Int. J. Semantic Comput.*, 17(4):545–568.

Cuzzocrea, A., Furfaro, F., Greco, S., Masciari, E., Mazzeo, G. M., and Saccà, D. (2005). A distributed system for answering range queries on sensor network data. In *3rd IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2005 Workshops), 8-12 March 2005, Kauai Island, HI, USA*, pages 369–373. IEEE Computer Society.

Cuzzocrea, A., Furfaro, F., and Saccà, D. (2003). Hand-olap: A system for delivering OLAP services on handheld devices. In *6th International Symposium on Autonomous Decentralized Systems (ISADS 2003), 9-11 April 2003, Pisa, Italy*, pages 80–87. IEEE Computer Society.

Cuzzocrea, A. and Mansmann, S. (2009). OLAP visualization: models, issues, and techniques. In *Encyclopedia of Data Warehousing and Mining, Second Edition (4 Volumes)*, pages 1439–1446. IGI Global.

Devastator, T. (2023). Mental health disorder.

Honda, K., Oda, T., Tanaka, D., and Notsu, A. (2015). A collaborative framework for privacy preserving fuzzy co-clustering of vertically distributed cooccurrence matrices. *Advances in Fuzzy Systems*, 2015:art. 729072.

Langone, R., Cuzzocrea, A., and Skantzos, N. (2020). Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *Data Knowl. Eng.*, 130:101850.

Organization, W. H. (2023). Cancer incidence.

Ouazzani, Z. E., Braeken, A., and Bakkali, H. E. (2021). Proximity measurement for hierarchical categorical attributes in big data. *Secur. Commun. Networks*, 2021:6612923:1–6612923:17.

Ram Mohan Rao, P., Murali Krishna, S., and Siva Kumar, A. (2018). Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data*, 5(1):33.

Russom, P. (2011). Big data analytics. *TDWI Best Practices report, Fourth Quarter*, 19(4):1–34.

Singh, A. K. and Kumar, J. (2023). A privacy-preserving multidimensional data aggregation scheme with secure query processing for smart grid. *J. Supercomput.*, 79(4):3750–3770.

Tran, H.-Y. and Hu, J. (2019). Privacy-preserving big data analytics a comprehensive survey. *Journal of Parallel and Distributed Computing*, 134:207–218.

Tsai, C.-W., Lai, C.-F., Chao, H.-C., and Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big data*, 2:1–32.

Wang, J., Fang, S., Liu, C., Qin, J., Li, X., and Shi, Z. (2020). Top-k closed co-occurrence patterns mining with differential privacy over multiple streams. *Future Gener. Comput. Syst.*, 111:339–351.

Wang, S., Sinnott, R., and Nepal, S. (2018). Pairs: Privacy-aware identification and recommendation of spatio-friends. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 920–931.

Wei, Y., Jia, J., Wu, Y., Hu, C., Dong, C., Liu, Z., Chen, X., Peng, Y., and Wang, S. (2024). Distributed differential privacy via shuffling versus aggregation: A curious study. *IEEE Trans. Inf. Forensics Secur.*, 19:2501–2516.

Wu, Y., Weng, D., Deng, Z., Bao, J., Xu, M., Wang, Z., Zheng, Y., Ding, Z., and Chen, W. (2021). Towards better detection and analysis of massive spatiotemporal co-occurrence patterns. *IEEE Trans. Intell. Transp. Syst.*, 22(6):3387–3402.