Empowering Multidimensional Machine Learning over Cloud-Enabled Big Data Infrastructures with *ClustCube*

Alfredo Cuzzocrea[1,2,*][a], Carmine Gallo[1][b] and Marco Antonio Mastratisi[3]
[1]*IDEA Lab, University of Calabria, Rende, Italy*
[2]*Department of Computer Science, University of Paris City, Paris, France*
[3]*SMARTCHAIN – ICT Technologies, Crotone, Italy*

Keywords: Big Data, Big Data Analytics, Multidimensional Machine Learning, Cloud-Enabled Big Data Infrastructures.

Abstract: *Multidimensional Machine Learning* is emerging as one of the key features in the whole *Big Data Analytics* landscape. Within this broad context, the OLAP paradigm is a reference pillar, and it represents the theoretical and methodological foundation of the so-called *Multidimensional Big Data Analytics* trend, an emerging trend in the Big Data era. In this paper, we show how the state-of-the-art *ClustCube* framework, which predicates the marriage between OLAP and *Clustering methodologies*, can be successfully used and exploited for effectively and efficiently supporting Multidimensional Big Data Analytics in real-life big data applications and systems.

1 INTRODUCTION

Big Data Analytics (e.g., (Russom, 2011; Zakir, *et al.*, 2015)) has imposed itself as one of the disruptive data technologies of the last decades. Indeed, the number of applications scenarios where Big Data Analytics has a relevant impact is paramount. Within this wide and various context, *Multidimensional Machine Learning* (e.g., (Orphanidou & Wong, 2017; Babanezhad, *et al.*, 2020; Moris, *et al.*, 2022; Zhong, *et al.*, 2020)) is emerging as one of the key features in the whole Big Data Analytics landscape. The main idea here consists in applying well-consolidated ML methodologies and methods to the core layer of well-understood big data analytics tasks directly, with the advantage of improving expressive power of analytical models and accuracy of retrieved results.

Within this broad context, the OLAP paradigm (Gray, *et al.*, 1997) is a reference pillar, and it represents the theoretical and methodological foundation of the so-called *Multidimensional Big Data Analytics* trend, an emerging trend in the Big Data era (Cuzzocrea, 2020; Cuzzocrea, 2021; Cuzzocrea 2022). The main assertion of Multidimensional Big Data Analytics theory relies in the fact that real-life datasets, and, especially, big datasets, are *inherently-multidimensional in nature* (e.g., (Iglesias, *et al.*, 2019)).

In this paper, we show how the state-of-the-art *ClustCube* framework (Cuzzocrea, 2015), which predicates the marriage between OLAP and *Clustering methodologies*, can be successfully used and exploited for effectively and efficiently supporting Multidimensional Big Data Analytics in real-life big data applications and systems. More into details, we devise and assess via implementations and case studies *an innovative ClustCube methodology in the context of Multidimensional Big Data Analytics*, and clearly prove its feasibility and reliability in real-life tools.

ClustCube represents a *cutting-edge* solution for integrating data from heterogeneous sources which vary both in content and format. This approach facilitates intelligent analysis and predictive processing in several sectors. It is important to note that data extracted from a distributed database can take significantly different forms than the original stored information. This diversity may be particularly evident in complex objects which could be generated

[a] https://orcid.org/0000-0002-7104-6415
[b] https://orcid.org/0009-0006-2637-5398

[*] This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France.

Cuzzocrea, A., Gallo, C. and Mastratisi, M.
Empowering Multidimensional Machine Learning over Cloud- Enabled Big Data Infrastructures with ClustCube.
DOI: 10.5220/0012727100003690
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 389-396
ISBN: 978-989-758-692-7; ISSN: 2184-4992
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

by complex *SQL* instructions involving multiple *JOIN* queries on distributed relational tables.

ClustCube model combines the power of clustering techniques on complex database objects with the versatility of OLAP in supporting multidimensional analysis and knowledge discovery from grouped complex database objects with mining opportunities and expressive power impossible for traditional methodologies. In *ClustCube* model, data cubes complex database objects are stored within data cube cells, rather than conventional SQL-based aggregations as in standard *Business Intelligence-oriented* OLAP data cubes.

On the other hand, the *ClustCube* model opens the door to innovative methods for supporting machine learning data analytics, by engrafting advanced methodologies such as *Multidimensional Clustering* and *Multidimensional Regression*.

The scheme of *ClustCube* is depicted in Figure 1, where different layers of the framework are shown.

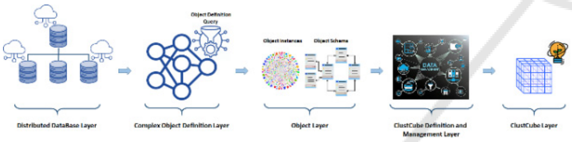
Object Definition Query
Object Instances
Object Schema
DATA
Distributed DataBase Layer
Complex Object Definition Layer
Object Layer
ClustCube Definition and Management Layer
ClustCube Layer

Figure 1: *ClustCube* Model.

2 RELATED WORK

The investigated research area contains several research proposals that are close to our work. In this Section, we report on some of the most noticeable of them.

In (Wu, *et al.*, 2023), authors discuss the challenge of *benchmarking for Cloud environments*, by noting that traditional benchmarks for OLAP are typically designed for *on-premise setups*. In response, they introduce *Raven*, a *Cloud-oriented OLAP benchmark* with a versatile architecture and varied workloads. *Raven* is tailored for assessing OLAP engine performance on Cloud platforms, boasting features like flexible architecture, diverse workloads, support for Cloud service deployment, integration with various Cloud OLAP engines (such as *Presto*, *SparkSQL*, *Kylin* and *Athena*) and the ability to evaluate different engine configurations. In essence, *Raven* offers users a comprehensive and adaptable tool for evaluating and understanding OLAP engine performance in Cloud environments by empowering them to make informed decisions about selecting and configuring OLAP services to meet their specific requirements.

Authors in (Kuschewski & Leis, 2021) discuss the challenge posed by the wide range of hardware instances offered by public cloud providers for *Cloud-based analytical query processing systems*. They propose the use of *white-box models* that take into account workloads, hardware and costs to determine the optimal instance configuration. According to the authors, this approach can guide the evolution of native Cloud OLAP systems. They argue that *black-box* models are unsuitable because they provide a few details and model-specific systems. Authors present a white-box model designed for OLAP workloads, along with an interactive tool to explore the model's predictions. This model includes *instance-local caching* and *materialization*, as well as imperfect scalability and network data exchange required for describing distributed processing.

In (Ribeiro, *et al.*, 2020), authors focus on OLAP analysis and parallel query processing in Cloud computing environments via using *C-ParGRES*. *C-ParGRES* is a system designed to execute OLAP queries in parallel by leveraging the resources available in a cloud computing environment. The main goal of this research is to enhance the performance and efficiency in processing OLAP queries, enabling better scalability and resource management in Cloud environments. Authors notice that several companies have moved their data to the Cloud using the concept of *Database as a Service* (DBaaS). Transferring databases to the Cloud poses various challenges related to flexible and scalable data management. While some of these companies have transitioned to *NoSQL databases*, many still rely on relational databases in the cloud to manage data, especially those critical for *decision-making* processes. With these considerations in mind, *C-ParGRES* explores database replication, *interquery* and *intraquery* parallelism to efficiently support OLAP queries in the cloud. *C-ParGRES* leverages cloud capabilities such as *on-demand* resource provisioning and elasticity. Additionally, *C-ParGRES* can create multiple independent virtual clusters for different databases and users.

Authors in (Zhan, *et al.*, 2019) analyze the phenomenon of the enormous growth of data in terms of size and variety. OLAP databases are assuming an increasingly crucial role in conducting *real-time analysis* with reduced latencies (for example, in the order of hundreds of milliseconds), especially when incoming requests are complex in nature. Additionally, these systems are expected to support significant query concurrency and high write throughput, as well as being able to handle queries on structured and complex data such as *JSON*, *vectors* and *texts*. In line with this, authors present *AnalyticDB*, a *real-time* OLAP database system developed at *Alibaba*. *AnalyticDB* maintains indexes of all columns asynchronously with an acceptable impact, thus ensuring *low latencies* for complex and spontaneous queries. Its storage engine extends the traditional *hybrid row-column layout* to allow for rapid retrieval of both structured and complex data. To manage *large-scale data* with high query concurrency and write throughput, *AnalyticDB* separates access paths for *reading* and *writing*. To further reduce query latencies, a new *storage-aware SQL optimizer and execution engine* has been developed to fully exploit the advantages of underlying storage and indexes. *AnalyticDB* has been successfully deployed on the *Alibaba Cloud* to meet the needs of numerous customers, both large and small.

In (Khrouf, *et al.*, 2018), authors focus the attention on documents containing relevant information for *decision-making processes*. Consequently, their analysis can assist decision-makers in better understanding their organizations' business processes and making well-founded decisions that cannot be achieved through traditional data warehouse inspection. However, it emerges that most textual data are not integrated into decision-making systems or considered during decision-making processes. This is due to various reasons, such as the lack of specific functionalities for document data warehouses. Therefore, to address this issue, it has become essential to incorporate textual data into *Decision Support Systems*. In this context, several attempts have been made, with most focusing on utilizing information stored in documents by employing a set of aspects to model documents according to different *user-proposed perspectives* and to enhance *OLAP over documents*. Authors identify two main categories for document storage and OLAP. The first one adopts the traditional multidimensional model by integrating specific extensions for textual data processing. The second one proposes specific models for document OLAP such as the *galaxy model* and the *diamond model*. In their research, authors present a new model called *CobWeb*, which extends the galaxy model for document OLAP. This model employs the concept of aspects, which are standard. An aspect is considered a viewpoint as it groups a set of data describing similar documents. By transforming each *aspect* into a *dimension* in the *CobWeb* document storage model, these aspects are used as possible *analysis axes*. In *CobWeb*, authors introduce several *significant extensions*, such as the constraint exclusion between dimensions, the ability to define recursive parameters, duplicate dimensions and correlated dimensions. Also, in *CobWeb*, authors introduce *four operators* for visualizing OLAP query results via using the concept of *Tag Cloud* as to assist *decision-makers* in interpreting their query results more effectively. These operators are: *TableCloudTags*, *Filter_Tag*, *CellCloudTags* and *Agg_Tag*.

Authors analyze *Cloud Data Warehouses* in (Dkaich, *et al.*, 2017), with their enormous size and high resource consumption that, very often, representing a significant burden for local infrastructures. *Cloud-based solutions* for data warehousing emerge as a promising answer to manage the immense amounts of involved data. Many companies extensively use data warehouses for data analysis, leveraging *XML* not only to handle *semi-structured* data but also to capitalize on the Web environment. The idea of integrating both solutions within a *parallel environment* therefore appears as a logical step. Authors propose leveraging XML not only for data storage and exchange but also to connect it to the distributed processing of multidimensional data. The study addresses the challenge of storing documents in distributed environments such as Cloud nodes, by exploring the possibility of combining data storage and decision analysis based on OLAP data cubes within Cloud environments via using the *MapReduce* model for query processing.

In (Dehne, *et al.*, 2015), authors introduce the problem of requests for *Online Transaction Processing* (OLTP) transactions in a data center, being these transactions typically accessing only *limited parts* of a database. On the other hand, OLAP queries need to aggregate vast portions of data, often causing performance issues. In reply to this challenge, (Dehne, *et al.*, 2015) presents *CR-OLAP*, an innovative *real-time OLAP system based on Cloud* that is built on a new distributed index structure for OLAP, namely the *distributed PDCR tree*. *CR-OLAP* leverages a scalable Cloud infrastructure consisting of multiple standard servers (*processors*). In other words, as the database size increases, *CR-OLAP* dynamically increases the number of processors to ensure optimal performance. In addition to this, the *distributed data structure PDCR tree* supports various *dimensional hierarchies* and efficiently processes complex dimension hierarchies, which are a fundamental element in OLAP systems. This system proves to be particularly efficient in handling complex OLAP queries that require aggregating large segments of the data warehouse, such as: reporting total sales for all stores located in California and New York during the months from February to May of all years. By conducting an evaluation of CR-OLAP on *Amazon EC2 Cloud* using the *TPC-DS benchmark dataset*, results demonstrate that CR-OLAP effectively scales with an increase in the number of processors, even for most complex queries. Derived results provide a response time of less than 0.3 seconds, considered *real-time*.

Finally, authors in (Cruz Lopes, *et al.*, 2014) explore mechanisms for handling transactional queries on *encrypted data*, which is slightly related to our research. However, little attention has been paid to understanding how a *Cloud-Hosted Data Warehouse* (CDW) should be encrypted to support query analysis. Since data is stored in the *DAS provider*, there are potential risks for sensitive data, such as financial information or medical records stored on an unreliable host. Therefore, directly executing queries on encrypted data can significantly improve query performance while preserving *data privacy*. With this goal in mind, authors introduce the concept of *direct query execution on encrypted data*, by illustrating the potential to improve query performance and preserve data privacy. They examine an approach to encrypt and query a CDW. Performance tests were conducted on the OLAP system developed by the authors, specifically tailored to the proposed encryption approach, to evaluate the system effectiveness in query processing. Results indicate that the overhead introduced by the proposed encryption approach decreases as the proposed system scales, as comparable to an unencrypted dataset. Furthermore, executing aggregations and data grouping directly on encrypted data stored in the server led to performance improvements (from 84.67% to 93.95%), when compared to the performance obtained by executing the same computational pattern on the client after decryption.

3 CLUSTCUBE FOR MULTIDIMENSIONAL BIG DATA ANALYTICS

In the context of *Multidimensional Big Data Analytics*, the aspect of *multidimensional cuboids* plays a fundamental role. These cuboids, each of which represents an entity in the multidimensional space of data provide a structured and comprehensive view of the data itself. It is important to emphasize that cuboids can vary significantly depending on the dimensions chosen and the level of detail applied to each of them.

The use of cuboids allows for a more in-depth analysis and understanding of the data by enabling the identification of complex relationships and patterns within them. This feature offers enormous potential to further optimize the entire process related to the creation, selection and management of cuboids. Starting from the *Cuboid Lattice Schema*, it is possible to significantly improve the phases involved in *Data Mart* development. This process includes several crucial stages:

- *Analysis and Reconciliation of Data Sources:* in this phase it is essential to carefully examine the available data sources by ensuring consistency and integrity before proceeding further.
- *Requirements Analysis:* fully understanding the needs and requirements of the system is fundamental by ensuring all end-user needs are met.
- *Conceptual Design:* key concepts and data relationships are outlined here creating a conceptual structure that will serve as a foundation for subsequent phases.
- *Loading and Validation of the Conceptual Schema:* this phase involves the practical implementation of the conceptual schema, verifying its effectiveness and consistency with the initial requirements.
- *Logical Design:* conceptual concepts translate into a logical model by defining tables, relationships and necessary *primary/foreign keys* for system implementation.
- *Physical Design:* technical and infrastructural aspects of system implementation are defined here, optimizing performance and scalability.
- *Feeding Design:* this phase focuses on defining processes for feeding data into the system by ensuring an efficient and reliable flow.

Utilizing the power of Big Data enables these phases to be addressed more efficiently and effectively. The *ClustCube methodology*, applied in this context is outlined in Figure 2, providing a visual and structured guide for implementing this innovative approach to data management.

ClustCube methodology is divided into seven phases or levels as shown in Figure 2. Below, we provide a detailed description of the operations carried out at each level, by using a specific data cube, called *TourismDC*, which is a tourism-sector data cube that will be described deeply in Section 4.

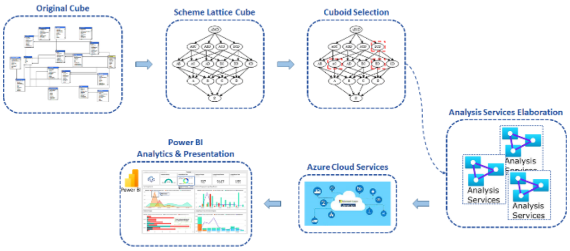

Figure 2: *ClustCube* Methodology for Multidimensional Big Data Analytics.

Level one represents the starting point. At this level we can observe the presence of the initial data cube, *TourismDC*. *TourismDC* consists of the *fact table* named *Reservation* to which sixteen dimensions are connected, expressing the analysis dimensions of the model. Among the main dimensions we observe *Accommodation*, *Point of Interest*, *Car Rental*, *Flight*, *Ferry*, *Taxi*, *Tour* and *Tourist*.

Second level implies the configuration of the Cuboid Lattice Schema relating to a data cube. Due to considerations regarding graphical representation, the displayed Cuboid Lattice Schema diagram actually corresponds to a *four-dimensional* data cube rather than the data cube *TourismDC*. This is because the Cuboid Lattice Schema of the data cube *TourismDC* would be too extensive to be displayed within such a schema.

However, the main issue concerns the intrinsic concept of the Cuboid Lattice Schema, which involves representing, in general, the concept that starting from a *fact table* connected to *n dimension tables*, it is possible to create a lattice of cuboids each of which represents a different degree of consolidation along one or more dimensions. In this context, therefore, level two of the schema indicates the representation of cuboids derived from the main data cube *TourismDC*.

At the third level the cuboid grid is still present, but three cuboids highlighted in red are noticeable indicating they have been selected. This highlighting reflects the selection process carried out by *ClustCube*, which can select cuboids belonging to different levels. This selection enables us to identify cuboids deemed significantly relevant for the subsequent analysis phase of the grid. Thanks to this targeted selection, we carefully choose cuboids capable of providing optimal results for our objective.

Level four presents the *Analysis Service Elaboration* service. Through the Analysis Service Elaboration operation, it is possible to process Analysis Services elements such as *tabular models*, *data cubes*, *dimensions* and *data mining models*. Multiple elements can be processed simultaneously, either sequentially or concurrently. If there are no elements requiring a specific processing order, concurrent processing can expedite the process. In the case of concurrent processing of elements, the operation can be configured to automatically determine the number of elements to process concurrently or manually specify this number. Sometimes, when processing analysis elements it is necessary to process elements dependent on them as well. *Analysis Services Processing* operation offers an option to process all dependent elements in addition to the selected ones. Processing performed on selected cuboids provides crucial information on both clusters and *Multidimensional Regression*.

Level five is represented by the *Cloud*. The *Computational Framework of Multidimensional Data Analytics System Support* relies on *Microsoft Azure Cloud* infrastructure. Crucial phases of the complete implementation of *Big Data Analytics* techniques are outlined in the following points:

- Analysis (*Big Data Mining*):
- Prediction (*Big Data Prediction*);
- Visualization (*Big Data Visualization*).

Peculiarity of each phase lies in the adoption of multidimensional modeling methodologies of Big Data aimed at enhancing the effectiveness and expressive power of all phases within the overall process of Big Data Analytics. The latter further demonstrates the flexibility of our methodology.

Level six and level seven focus on the visualization and processing capabilities, as well as prediction, of the *PowerBI* component. The framework introduces suitable solutions for big data visualization that utilize advanced graphical tools for accessing Big Data, as to really support and guide the big data analytics phase via simplifying data understanding and interpretation.

In particular, the types of visualization components that the framework aims to support are as follows:

- *Built-In Big Data Visualization Components:* these visualization components create classic graphical objects (histograms, scatter plots, pies, etc.) directly from the knowledge extracted from Big Data.
- *User-Defined Data Visualization Components:* these visualization components are graphical objects whose appearance can be defined based on specific analysis goals, either natively or by composing the basic Built-In Big Data Visualization Components.

4 A PROOF-OF-CONCEPT FOR CLUSTCUBE

In this Section, we provide a detailed proof-of-concept of the capabilities of *ClustCube* in a real-life setting represented by the tourism sector. We first introduce the data cube *TourismDC*, the central entity of our proof-of-concept. Then, interesting multidimensional cuboids can be derived from the main data cube.

Conceptual diagram in Figure 3 illustrates the *Dimensional Fact Model* (DFM), through which it is feasible to represent data within the OLAP data cube *TourismDC*. It highlights the dimensions of analysis and the relevant facts related to the representation of the modeled reality.

Fundamental aspects of the model are:

- providing support for *conceptual design*;
- creating an environment where users can query intuitively and guided.

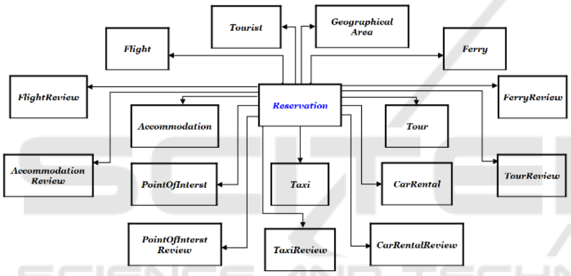

Figure 3: DFM of the Data Cube *TourismDC*.

Graphical representation of the DFM assumes the form of a *Star Schema*. At the center of the schema lies the *Reservation* table, serving as the *fact table* since it contains various measures enabling significant results. Furthermore, it holds considerable importance in the *decision-making* process. The *Reservation* table contains all bookings made by tourists for available services.

Starting from the DFM, the data cube *TourismDC* has been implemented through the integrated environment of *Microsoft Visual Studio 2019* (see Figure 4).

This implementation allows us to realize both the big multidimensional data management and analytics procedures, in order to effectively test the properties and the potentialities of the proposed framework in real-life settings.

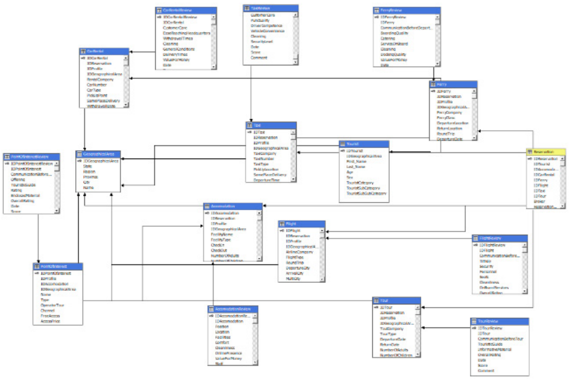

Figure 4: Data Cube *TourismDC*.

Here, the *fact table* is associated with sixteen *dimensions* and their corresponding relationships that delineate the various analytical dimensions of the model. These dimensions include *Accommodation*, *Point of Interest*, *Car Rental*, *Flight*, *Ferry*, *Taxi*, *Tour*, *Tourist* and *Geographical Area*. Some of these dimensions feature *hierarchies* to allow for a more detailed analysis. In our context, the dimensions represent the services available to tourists. Other dimensions include: *Accommodation Review*, *Car Rental Review*, *Flight Review*, *Ferry Review*, *Tour Review*, *Taxi Review* and *Point of Interest Review*. The latter dimensions listed, as suggested by their names, represent the evaluations expressed by tourists regarding the quality of the services used.

As regards the multidimensional big data analytics properties of the framework, Figure 5 illustrates a case of *Multidimensional Clustering Analysis*.

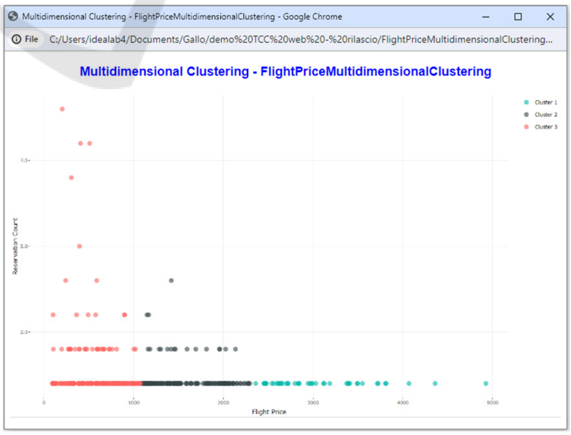

Figure 5: Section *Multidimensional Clustering Analysis*.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have provided our methodology for applying the state-of-the-art *ClustCube* framework to the real-life problem of supporting Multidimensional Machine Learning over Cloud-enabled big data infrastructures, and we have shown its proof-of-concept in the tourism sector. This model can be further extended, for instance via supporting the discovery of new *communities* among tourists (e.g., (Wu, *et al.*, 2013; Chen, et al., 2021)).

Future work is mainly oriented to engraft into our framework innovative aspects of big data processing (e.g., (Cuzzocrea & Mansmann, 2009; Islam, *et al.*, 2017; Langone, *et al.*, 2020; Barkwell, *et al.*, 2018; Bobek, *et al.*, 2022; Hiremath, *et al.*, 2023)).

ACKNOWLEDGEMENTS

This work was partially 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

Babanezhad, M., Marjani, A., & Shirazian, S. (2020). Multidimensional Machine Learning Algorithms to Learn Liquid Velocity Inside a Cylindrical Bubble Column Reactor. *Scientific Reports 10(1)*, art. 21502.

Barkwell, K.E., Cuzzocrea, A., Leung, C.K., Ocran, A.A., Sanderson, J.M., Stewart, J.A., & Wodi, B.H. (2018). Big Data Visualisation and Visual Analytics for Music Data Mining. In: *IV'18, 22nd International Conference Information Visualisation*, pp. 235-240.

Bobek, S., Kuk, M., Szelazek, M., & Nalepa, G.J. (2022). Enhancing Cluster Analysis with Explainable AI and Multidimensional Cluster Prototypes. *IEEE Access 10*, pp. 101556-101574.

Chen, Y., Chen, R., Hou, J., Hou, M., & Xie, X. (2021). Research on Users' Participation Mechanisms in Virtual Tourism Communities by Bayesian Network. *Knowledge Based Systems 226*, art. 107161.

Cruz Lopes, C., Cesário Times, V., Matwin, S., Rodrigues Ciferri, R., & Dutra de Aguiar Ciferri, C. (2014). Processing OLAP Queries over an Encrypted Data Warehouse Stored in the Cloud. In: *DaWaK'14, International Conference on Big Data Analytics and Knowledge Discovery*, pp. 195-207.

Cuzzocrea A., & Mansmann S. (2009). OLAP Visualization: Models, Issues, and Techniques. *Encyclopedia of Data Warehousing and Mining*, pp. 1439-1446.

Cuzzocrea, A. (2015) Computing and Mining ClustCube Cubes Efficiently. In: *PAKDD'15, 19th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 146-161.

Cuzzocrea, A. (2020). OLAPing Big Social Data: Multidimensional Big Data Analytics over Big Social Data Repositories. In: *ICCBDC'20, 4th International Conference on Cloud and Big Data Computing*, pp. 15-19.

Cuzzocrea, A. (2021). Innovative Paradigms for Supporting Privacy-Preserving Multidimensional Big Healthcare Data Management and Analytics: The Case of the EU H2020 QUALITOP Research Project. In: *SWH@ISWC'21, 4th International Workshop on Semantic Web Meets Health Data Management*, pp. 1-7.

Cuzzocrea, A. (2022). Multidimensional Big Data Analytics over Big Web Knowledge Bases: Models, Issues, Research Trends, and a Reference Architecture. In: *BigMM'22, 8th IEEE International Conference on Multimedia Big Data*, pp. 1-6.

Dehne, F.K.H.A., Kong, Q., Rau-Chaplin, A., Zaboli. H., & Zhou, R. (2015). Scalable real-time OLAP on cloud architectures. *Journal of Parallel and Distributed Computing 79-80*, pp. 31-41.

Dkaich, R., El Azami, I., & Mouloudi, A. (2017). XML OLAP Cube in the Cloud towards the DWaaS. *International Journal of Cloud Computing 7(1)*, pp. 47-56.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., & Pirahesh, H. (1997). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Mining and Knowledge Discovery 1(1)*, pp. 29-53.

Hiremath, S., Shetty, E., Prakash, A.J., Sahoo, S.P., Patro, K.K., Rajesh, K.N., & Pławiak, P. (2023). A New Approach to Data Analysis Using Machine Learning for Cybersecurity. *Big Data and Cognitive Computing, 7(4)*, art. 176.

Iglesias, F., Zseby, T., Ferreira, D., & Zimek, A. (2019). MDCGen: Multidimensional Dataset Generator for Clustering. *Journal of Classification 36*, pp. 599-618.

Islam, O., Alfakeeh, A., & Nadeem, F. (2017). A Framework for Effective Big Data Analytics for Decision Support Systems. *International Journal of Computer Networks and Applications 4(5)*, pp. 129-137.

Khrouf, O., Khrouf, K., & Feki, J. (2018). CobWeb Multidimensional Model and Tag-Cloud Operators for OLAP of Documents. *International Journal of Green Computing 9(2)*, pp. 46-48.

Kuschewski, M., & Leis, V. (2021). White-Box OLAP Performance Modelling for the Cloud. In: *CIDR'21, Conference on Innovative Data Systems Research*.

Langone, R., Cuzzocrea, A., & Skantzos, N. (2020). Interpretable Anomaly Prediction: Predicting Anomalous Behavior in Industry 4.0 Settings via

Regularized Logistic Regression Tools. *Data & Knowledge Engineering 130*, art. 101850.

Moris, D., Henao, R., Hensman, H., Stempora, L., Chasse, S., Schobel, S., & Elster, E. (2022). Multidimensional Machine Learning Models Predicting Outcomes after Trauma. *Surgery 172(6)*, pp. 1851-1859.

Orphanidou, C., & Wong, D. (2017). Machine Learning Models for Multidimensional Clinical Data. *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, pp. 177-216.

Ribeiro, M.W.M., Lima, A.A.B., & De Oliveira, D. (2020). OLAP Parallel Query Processing in Clouds with C-ParGRES. *Concurrency and Computation: Practice and Experience 32(7)*, art. e5590.

Russom, P. (2011). Big Data Analytics. *TDWI best practices report, fourth quarter, 19(4)*, pp. 1-34.

Wu, T., Gu, R., Li, Y., Ma, H., Chen, Y., Zhu, Y., Yu, X., Xu, T., & Huang, Y. (2023). Raven: Benchmarking Monetary Expense and Query Efficiency of OLAP Engines on the Cloud. In: *DASFAA'23, 29th International Conference on Database Systems for Advanced Applications*, pp. 593-605.

Wu, Z., Yin, W., Cao, J., Xu, G., & Cuzzocrea, A. (2013). Community Detection in Multi-Relational Social Networks. In: *WISE'13, 14th International Conference on Web Information Systems Engineering*, pp. 43-56.

Zakir, J., Seymour, T., & Berg, K. (2015). Big Data Analytics. *Issues in Information Systems 16(2)*, pp. 81-90.

Zhan, C., Su, M., Wei, C., Peng, X., Lin, L., Wang, S., Chen, Z., Li, F., Pan, Y., Zheng, F., & Chai, C. (2019). AnalyticDB: Real-Time OLAP Database System at Alibaba Cloud. *Proceedings of VLDB Endowment 12(12)*, pp. 2059-2070.

Zhong, X., Luo, T., Deng, L., Liu, P., Hu, K., Lu, D., & Zheng, H. (2020). Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study. *JMIR Medical Informatics 8(11)*, art. e19069.