

# QFLS: A Cloud-Based Framework for Supporting Big Healthcare Data Management and Analytics from Big Data Lakes: Definitions, Requirements, Models and Techniques

Alfredo Cuzzocrea<sup>1,2</sup> <sup>a</sup> and Selim Soufargi<sup>1</sup> <sup>b</sup>

<sup>1</sup>*iDEA Lab, University of Calabria, Rende, Italy*

<sup>2</sup>*Department of Computer Science, University of Paris City, Paris, France*

**Keywords:** Big Healthcare Data, Big Data Management, Big Data Analytics, Cloud-Based Frameworks.

**Abstract:** This paper introduces definitions, requirements, models and techniques of *QUALITOP Federated Big Data Analytics Learning System* (QFLS), a Cloud-based framework for *supporting big healthcare data management and analytics from big data lakes*. QFLS anatomy and main functionalities are described, along with the main software solutions proposed with the framework.


## 1 INTRODUCTION


Nowadays, *big data* management and analytics play a disruptive role in modern database research (e.g., (Chen et al., 2013; Zhang et al., 2015; Chaudhuri, 2012; Cuzzocrea et al., 2014; Campan et al., 2017; Balbin et al., 2020)), with a special touch on both theory and systems. When applied to *healthcare domains* (e.g., (Dash et al., 2019; Sun and Reddy, 2013; Patil and Seshadri, 2014)), relevant challenges arise, starting from big data representation to big data indexing, from big data understanding to big data analytics, and so forth. These open issues have been recently highlighted by authoritative proposals in the scientific field (e.g., (Ghayvat et al., 2022; Ding et al., 2022; Parimanam et al., 2022)).

Indeed, healthcare domains dictate special requirements, among that the need for supporting advanced analytics while ensuring the *privacy* of input big healthcare datasets is a first-class point (e.g., (Onesimu et al., 2022; Abbasi and Mohammadi, 2022; Singh et al., 2022)). This is one of the main goal of the EU H2020 project QUALITOP (QUALITOP, 2023). QUALITOP aims at devising a collection of models, techniques and algorithms for supporting prediction and recommendation for the *quality of life* of cancer patients treated with the modern immunotherapy treatment (e.g., (Vinke et al., 2023; Beaulieu et al., 2022)).

One of the most distinctive characteristics of QUALITOP is represented by the so-called *QUALITOP big data lake*, a big data lake (e.g., (Cuzzocrea, 2021)) oriented to collect *anonymized* data/information from a reference *data federation* built at the medical premises, and to support big data analytics and big data predictive analytics over such datasets. This, with the final goal of supporting analysis and recommendations of immunotherapy-treated cancer patients during their post-trauma life. The big data lake relies within the general QUALITOP software platform (e.g., (Elgammal and Krämer, 2021)).

In order to fulfill the requirements super-imposed by such scenarios, we propose the *QUALITOP Federated Big Data Analytics Learning System* (QFLS), a Cloud-based federated framework for supporting big healthcare data management and analytics from big data lakes. Among various results, QFLS introduces the so-called *Tree-Like Analytical Query* (TLAQ) model, a powerful analytical model for healthcare analytics. Given a TLAQ  $Q$ , every node  $n \in Q$  is equipped by query node  $Q_n$ , modeled as follows:  $Q_n = \langle A_n, \Sigma_{A_n}, P_n \rangle$ , such that: (i)  $A_n$  is the target query attribute; (ii)  $\Sigma_{A_n}$  is a constraint over  $A_n$ ; (iii)  $P_n$  is an aggregate operator over  $A_n$ . Another relevant characteristic of the TLQA model is that, given two query nodes  $Q_n$  and  $Q_m$ , such that  $Q_n < Q_m$ , there not exists a strict hierarchical relation between  $Q_n$  and  $Q_m$ , meaning that  $A_m \not\subseteq A_n$ . This special feature is particularly suitable to support *precision medicine processes*. When executed, TLAQ generate a tree-like analytical

<sup>a</sup>  <https://orcid.org/0000-0002-7104-6415>

<sup>b</sup>  <https://orcid.org/0009-0000-5476-9403>

data structure that indexes anonymized datasets, made available for several privacy-preserving big data analytics purposes (e.g., (Lin et al., 2016)).

In more details, QFLS is a *data federation solution* that is based on the big data processing framework *Hadoop cluster*, to which all queries are dispatched. In addition, specialized *Apache-Spark-based servers* that run remotely within each of the data federation sources are introduced. QFLS is developed and deployed on top of the *iDEA Lab Cloud*, located at University of Calabria, Rende, Italy, which is composed of 21 VMs equipped with *Windows Server Essentials 2019 OS*, each having 32GB of memory, 8-Core CPUs, 32GB RAM and 60GB HDD.

It goes without saying that it is of best interest to run Spark jobs on the Hadoop cluster in a *MapReduce* manner, especially for larger datasets, and to be able to do that, fine tuning Spark job submission onto the cluster was necessary. Using full potential of the Hadoop cluster is a key to bring effectiveness and efficiency during query executions. Specifically, the server runs on Hadoop (through *YARN* resource manager) with fine-tuned worker instances, worker memory and CPU, and driver memory and CPU (e.g., (Chen et al., 2017; Gounaris and Torres, 2018)).

The latter sub-system represents the *QFLS Core* component of QFLS, which is the Cloud-based engine that runs everything. QFLS encompasses more two components, namely *QFLS-ADPT* and *QFLS-ADAT*. The *QUALITOP Anonymized Dataset Population Tool* (QFLS-ADPT) is a web-based tool for managing the anonymized dataset at the federation node. The *QUALITOP Anonymized Dataset Analytics Tool* (QFLS-ADAT) is a web-based tool for supporting big healthcare analytics and predictive analytics over the anonymized datasets stored in the federation, via the TLAQs.

## 2 QFLS ANATOMY

Figure 1 shows the main blueprint of the big data processing flow supported by QFLS. Here, we introduce an example scenario where the following two nodes occur: one QFLS federated node, located in France, and one QFLS Core node, located in Italy. Therefore, the French node provides (anonymized) healthcare data, and the Italian node support big data analytics and predictive analytics over these data.

As shown in Figure 1, at the French node (1), where the target healthcare dataset  $D$  is located, medical operators generate an anonymized version of  $D$ , denoted by  $D'$  (3), according to specific medical guidelines (2), yet compliant with the GDPR.

Then, the analytics tools located at the Italian node (4) execute aggregate queries, defined by the input TLAQ (8), which, in turn, is driven by target big data analytics tools defined within the QUALITOP big data lake, over the remote French node, via federated query algorithms based on Apache Spark, and the anonymized representation of the result is so-obtained (5). The final TLAQ analytics answer is shaped as a tree (9) such that each node indexed a proper anonymized healthcare dataset, empowered by the Cloud computing potentialities (e.g., distribution, indexing, load balancing, mirroring, and so forth). The latter data structure is accessed by medical decision makers (10) who provide the final big data analytics and predictive analytics (11).

It should be noted some relevant characteristics of QFLS, which we summarize as follows:

1. **Data Federation.** While QFLS Core is entirely Cloud-based, the main processing of QFLS happens over data federation nodes, according to severe data privacy protection guidelines. This scenario fully converges with real-life systems, where data federation participants are interested in participating to the federation (for data analytics purposes) but are not willing to unveil their data. The latter is the common case of medical centers.
2. **Privacy-Preserving Advanced Analytical Models.** QFLS encompasses the TLAQ analytical model, a powerful privacy-preserving analytical model particularly target to support precision medicine processes that, by definition, are built upon *lazy aggregations*. For instance, at first, physicians may be interested in analyzing COVID-19 data about female patients who live in Canada and whose age in between the range [25-50], then, *within* this range, they may want to analyze COVID-19 data about female people of age 30 who are resident in Toronto and have also been diagnosed with Tuberculosis.
3. **Big Data Analytics & Big Data Predictive Analytics.** QFLS not only supports big data analytics, mostly driven by the TLAQ model, but also the so-called big data predictive analytics. Indeed, retrieved TLAQ analytics can be combined in a *multidimensional fashion*, and so-derived summary data can be used to develop fortunate big multidimensional analytics metaphors (e.g., (Cuzzocrea et al., 2004)).
4. **Flexible Big Data Processing Tools Integration.** QFLS fully relies on the smoothly integration of several big data processing tools, mostly within the eco-system defined by Hadoop, such Apache

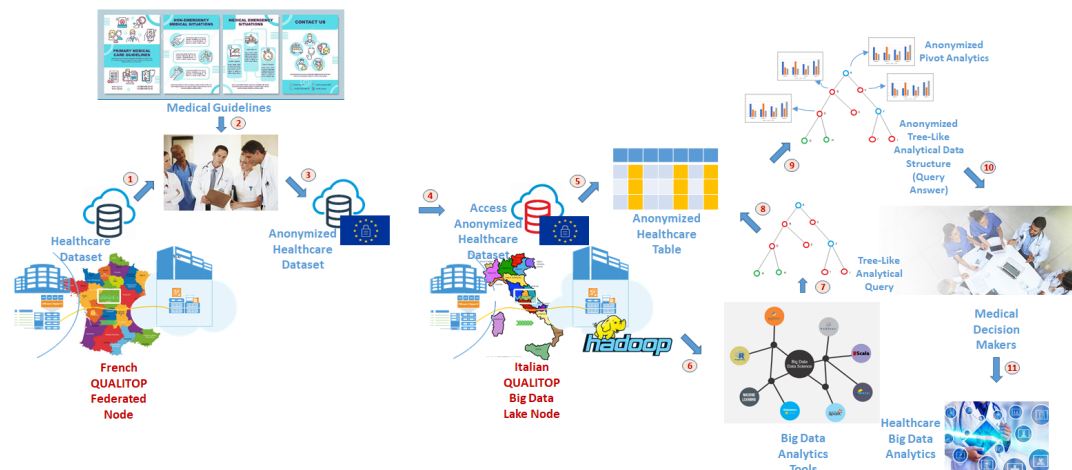


Figure 1: QFLS Main Blueprint.

Spark, YARN, Hive, MongoDB, etc. This ensures high data-availability, high scalability and, moreover, a full tendency to be further integrated within more complex big data stack architectures.

### 3 RELATED WORK

The problem of supporting big healthcare data management and analytics has been of great interest recently, even due to the COVID-19 outbreak (e.g., (Liu et al., 2023b; Dimitsaki et al., 2023)). In this Section, we provide a brief overview of most relevant proposals in this scientific area.

(Ghayvat et al., 2022) focuses on *Internet of Things (IoT)-based Healthcare services*, which are becoming more widespread today, continuously generate huge amounts of big data. Due to the data magnitude, data intricacy, privacy preservation, data integrity and identity verification requirements, indeed novel research challenges and issues in healthcare big data service management arise. To overcome these problems, author propose a *scalable computing system* that provides verifiable data access mechanism for IoT-enabled health data analytics in the big data ecosystem. There are two primary sub-architectures in the proposed architecture, namely a *big data analytics tracking system* and a *derived blockchain-based data storage/access system*. This approach leverages big data systems and blockchain architecture to analyze, and securely store data from IoT-enabled devices and allow verified access to the stored data. The zero-knowledge protocol is used to ensure that no information is accessible to unauthenticated users alongside avoiding data linkability. The results demonstrate the effectiveness of the proposed method

to solve the problems of big data analytics and privacy issues in healthcare.

(Ding et al., 2022) proposes two *differentially private algorithms*, i.e., *Output Perturbation with aGM* (OPERA) and *Gradient Perturbation with aGM* (GRPUA) for empirical risk minimization, a useful method to obtain a globally optimal classifier, by leveraging the analytic Gaussian mechanism (aGM) to achieve privacy preservation of sensitive medical data in a healthcare system. Authors theoretically analyze and prove utility upper bounds of proposed algorithms and compare them with prior algorithms in the literature. The analyses show that in the high privacy regime, the proposed algorithms can achieve a tighter utility bound for both settings: strongly convex and non-strongly convex loss functions. Besides, the proposed private algorithms are evaluated against five benchmark datasets. The simulation results demonstrate that these approaches can achieve higher accuracy and lower objective values compared with existing ones in all three datasets while providing differential privacy guarantees.

(Parimanam et al., 2022) notices that, current health information systems, when coupled big data trends, fail to maintain a highly organized analysis and processing of health data for analytics purposes. In addition, it collects issues that play an important role in determining *quality*. In this article, authors propose a *Hybrid Optimization based Learning technique for Multi-Disease analytics* (HOLMD) from healthcare big data using optimal pre-processing, clustering, and classifier. First, authors introduce a *capuchin* search based optimization algorithm for pre-processing which removes the unwanted artifacts to enhance the detection accuracy. Second, a modified *Harris Hawks Optimization based*

*Clustering* (MHHOC) technique is introduced used to select optimal features among multiple features which discovers the subgroups and reduce the dimensionality issues. Performance of proposed HOL-MD technique is evaluated using standard US healthcare-organization SUSY and HIGGS datasets, and it turns that existing state-of-art techniques are outperformed in terms of *accuracy*, *precision*, *recall* and *F-measure*, respectively.

## 4 QFLS: REQUIREMENTS AND MAIN FUNCTIONALITIES

This Section highlights, mostly from a software-engineering and component-architecture point-of-view, the requirements and the main functionalities of QFLS.

### 4.1 Data Ingestion and Storage: The QADPT Component

In order to enable data ingestion into QFLS, federated nodes are enabled with a client web component that ensures data can be referred by QFLS, and can be processed for later analytics, according to FAIR principles (*Findable*, *Accessible*, *Interoperable* and *Repeatable*).

In order to ingest the (anonymized) data, current QADPT provides an interface to select data from local file system as CSV files. The data are then stored into *Hive* table and QADPT notifies QFLS of the ingested data through updating the data setting of QFLS federated nodes. After updating the QFLS system configuration, users are allowed to see the added datasets in the respective node and use them for submitting their TLAQ (*federated query answering*). To note that, at ingestion time, data are cleansed and transformed to meet QFLS standards and constraints (date format, column values encoding issues, and so forth). On the other hand, it should be noted that the QFLS-based anonymization mechanism ensures that users cannot never access the original data, but only privacy-preserving summarized versions of them.

QADPT is also capable of removing a dataset from QFLS federated node or also list existing ones in a intuitive browsable interface. QADPT implements a *role-based access control*, and allows authentication of medical operator users solely.

QADPT is a client web component implemented using *Spring v. 5.3* and deployed over a *Tomcat v. 8.5* instance. Main functionalities of QADPT is to load ingested data into *Hive*. The insertion into *Hive*

is performed through a *JDBC* client which takes care of loading the dataset (then, a CSV file) into a corresponding table in *Hive*, it also, as mentioned, upsert the associated (dataset) information into a *HDFS*-based configuration system file (.ini) depending on whether the node is previously existent in QFLS (update) or the node is newly created (insert).

### 4.2 Data Analytics: The QADAT Component

In order to enable data scientists and medical operators with insightful recommendation on the healthcare data, QFLS provides QADAT, a client web component that offers numerous tools to be used, all in a matter of a deep analysis of data and to eventually provide recommendations, or, better, *predictive analytics*, upon these global analysis. Figure 2 shows the main workflow for predictive analytics supported by QFLS.

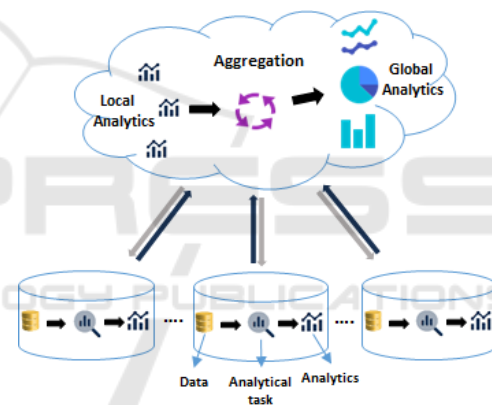


Figure 2: QFLS Workflow for Predictive Analytics.

In more details, Figure 2 shows the interaction between the Cloud-based QFLS framework QFLS Core and the federated nodes. At each federated node, data are processed as partial analytical tasks, via the TLAQ queries, and (partial) results are sent back to the Cloud for aggregation with other partial results similarly obtained from other nodes. The partial results are then processed and structured as needed and *global analytics* are then derived. QFLS produces from the TLAQ-shaped global analytics a *visual dashboard* consisting of predictive analytics tools. These can be in many forms: clustering, ROC areas, confusion matrices, and so forth. These predictive analytics tools serve for recommendation purposes by the competent medical operators for the patients. It is also worth noticing that the aforementioned QFLS system set-up preserves data locality which would increase anonymity and produce more tailored analytics. All of the data



processing logic is executed on the QFLS Core, in a MapReduce fashion.

QADAT is a client web component implemented using Spring v. 5.3 from which users can analyze the data stored in all linked federated nodes, in an anonymized manner. Like for the case of QADPT, the choice of Spring was made because it is compliant with the employed *JDK* version 8, and, with our set of technological solutions such as those used for the server (e.g., Spark) and for deployment (e.g., Tomcat), that was the most conforming solution among all the available ones. *JWT* authentication (Spring security) along with role based access were also part of our Spring application. Indeed, *data analyst* role has full components access whereas *medical operator* modules access is restricted. For instance, a medical operator user cannot use some SQL-based modules. This general idea could be further extended by adopting *adaptive* metaphors developed in the context of web information systems (e.g., (Cannataro et al., 2001; Cannataro et al., 2002)).

In order to test our Spring application, we used *JUnit* v. 5 to perform unit tests over the developed code base. In addition, deployment of each of the client applications was achieved thanks to Tomcat application server, like for QADPT.

Finally, given the need to store information about the user (such as their credentials) and other information related to the implemented functionalities, a *MongoDB* database was installed and used for the purpose.

#### 4.2.1 A Cloud-Based, Client-Server Solution

In order to reach the data at each of the federated nodes, QADAT employs Java *RMI* technology to transport result data (i.e., aggregations) from one machine to another. In addition, and since data could potentially be large, conveying them over the network requires the usage of specialized libraries such as *RMIIO* to enable large data to be transported over the network in a convenient way. More specifically, a Spark running instance server receives the remote method invocation request from the client through an *RMI* call, processes the dataset accordingly (in a MapReduce fashion) over the Hadoop cluster, and then returns the results to the client. The results are then gathered, structured and displayed to users.

### 4.3 Main Components and Functionalities

QFLS is intended to satisfy basic analytical needs by performing analytical computations over diverse data

sources in a federated context. First, QFLS enables data exploration throughout its *Data Federation discovery* (DFD). DFD exposes, based on the configuration file, the content of each of the federated node in terms of datasets. Second, the *Anonymized Dataset Analysis* (ADA) component provides an analysis of each of the target anonymized dataset, by providing details on its level of anonymization of the attributes, and other statistics such as the minimum value, the maximum value, the type. Third, the *Anonymized Analytics Environment* (AAE) allows users to submit TLAQ queries onto the remote target datasets and to retrieve analytics in a tree-like shaped analytical structure through. QFLS is also capable of deriving a dashboard over the query answers in order to enable recommendations based on the obtained analytics through the *Predictive Analytical Environment* (PAE). Furthermore, QFLS enables SQL queries to be submitted and executed on a target dataset, and to retrieve the corresponding result set in an intuitive interface, easily navigable by non-medical persons, through the *SQL Query Environment* (SQE). Each submitted tree-like analytical query is stored in a database. All of the aforementioned modules use Java *RMI* to request and retrieve results from the remote Spark-instance running server.

Figure 3 shows the *UML package diagram* of QFLS, as to highlight the inter-dependencies among the various components.

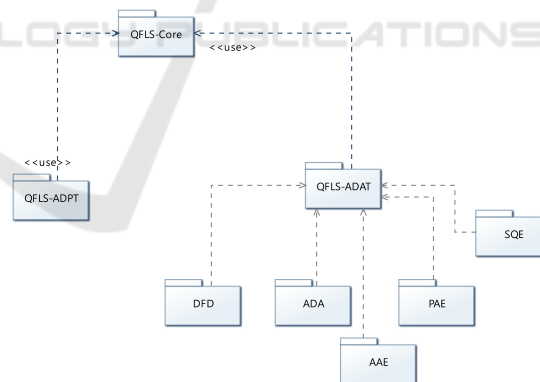


Figure 3: UML Package Diagram of QFLS.

Since, within the scope of the QUALITOP project, medical data at the federated nodes cannot be accessed, all of the data processing logic occurs on server side (on each federated node) and only aggregate information is returned back to the QADAT component. This ensures the necessary level of anonymization, as per the requirements of the QUALITOP project.

To better understand the data access constraint, we adopt the strategy of retrieving *distinct* values within

dataset columns, which are necessary at various sub-components of QADAT (e.g., AAE component). In this case, Spark processing yields only distinct values through adequately processing the anonymized dataset on the corresponding node, while data are never returned raw, nor are fully-processed at the client side. Indeed, even though data are anonymized on each federated node, data access may still expose certain sensitive information such as patient names or their diseases (e.g., (Sweeney, 2002)). In addition to this measure, identifier columns are removed at the time of processing the dataset at the federated node, as to reinforce anonymity of data and to provide optimal security guarantees over the privacy of patients and data owners.

Finally, QFLS is obviously extendable to a *federated machine learning system* (e.g., (Yang et al., 2019; Fu et al., 2022)), where the processing at each node level would be based on a machine learning model rather than a data aggregation model, thus resulting in a *global machine model*. Under this assumption, *SparkML* would be our best bet to employ for embedding further capabilities to our framework (e.g., (Omran et al., 2021; Mohamed et al., 2021)).

## 5 CONCLUSIONS AND FUTURE WORK

This paper has described in details definitions, requirements, models and techniques of QFLS, a Cloud-based framework for *supporting big healthcare data management and analytics from big data lakes*. QFLS anatomy and main functionalities have been described, along with the main software solutions proposed with the framework.

Concepts and guidelines deriving from the proposed framework also opens the door to emerging research challenges for the future. Among those, an interesting research line consists in extending our framework as to deal with advanced machine learning techniques, such as those experimented in other research efforts (e.g., (Cuzzocrea et al., 2017; Leung et al., 2019; Coronato and Cuzzocrea, 2022; Liu et al., 2023a; Adeoye et al., 2023; Tan et al., 2023)).

## ACKNOWLEDGEMENTS

This research has been funded by the EU H2020 QUALITOP research project - Call Reference: H2020 - SC1-DTH-01-2019; Project Number: 875171.

## REFERENCES

- Abbasi, A. and Mohammadi, B. (2022). A clustering-based anonymization approach for privacy-preserving in the healthcare cloud. *Concurr. Comput. Pract. Exp.*, 34(1).
- Adeoye, J., Koohi-Moghadam, M., Choi, S., Zheng, L., Lo, A. W. I., Tsang, R. K., Chow, V. L. Y., Akinshipo, A., Thomson, P., and Su, Y. (2023). Predicting oral cancer risk in patients with oral leukoplakia and oral lichenoid mucositis using machine learning. *J. Big Data*, 10(1):39.
- Balbin, P. P. F., Barker, J. C. R., Leung, C. K., Tran, M., Wall, R. P., and Cuzzocrea, A. (2020). Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science*, 176:3009–3018.
- Beaulieu, E., Spanjaart, A., Roes, A., Rachet, B., Dalle, S., Kersten, M. J., Maucourt-Boulch, D., and Jalali, M. S. (2022). Health-related quality of life in cancer immunotherapy: a systematic perspective, using causal loop diagrams. *Qual Life Res.*, 31(8).
- Campan, A., Cuzzocrea, A., and Truta, T. M. (2017). Fighting fake news spread in online social networks: Actual trends and future research directions. In *IEEE Big-Data, 2017*, pages 4453–4457. IEEE Computer Society.
- Cannataro, M., Cuzzocrea, A., and Pugliese, A. (2001). A probabilistic approach to model adaptive hypermedia systems. In *Proceedings of the First International Workshop on Web Dynamics, WebDyn@ICDT 2001, London, UK, January 3, 2001*, pages 50–60.
- Cannataro, M., Cuzzocrea, A., and Pugliese, A. (2002). XAHM: an adaptive hypermedia model based on XML. In *Proceedings of the 14th international conference on Software engineering and knowledge engineering, SEKE 2002, Ischia, Italy, July 15-19, 2002*, pages 627–634.
- Chaudhuri, S. (2012). What next?: a half-dozen data management research goals for big data and the cloud. In Benedikt, M., Krötzsch, M., and Lenzerini, M., editors, *ACM SIGMOD-SIGACT-SIGART PODS, 2012*, pages 1–4. ACM.
- Chen, D., Chen, H., Jiang, Z., and Zhao, Y. (2017). An adaptive memory tuning strategy with high performance for spark. *Int. J. Big Data Intell.*, 4(4):276–286.
- Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., and Zhou, X. (2013). Big data challenge: a data management perspective. *Frontiers Comput. Sci.*, 7(2):157–164.
- Coronato, A. and Cuzzocrea, A. (2022). An innovative risk assessment methodology for medical information systems. *IEEE Trans. Knowl. Data Eng.*, 34(7):3095–3110.
- Cuzzocrea, A. (2021). Big data lakes: Models, frameworks, and techniques. In *IEEE BigComp, 2021*, pages 1–4. IEEE.
- Cuzzocrea, A., Furfaro, F., Mazzeo, G. M., and Saccà, D. (2004). A grid framework for approximate aggregate

- query answering on summarized sensor network readings. In *OTM 2004 Workshops, 2004*, volume 3292, pages 144–153.
- Cuzzocrea, A., Leung, C. K., and MacKinnon, R. K. (2014). Mining constrained frequent itemsets from distributed uncertain data. *Future Gener. Comput. Syst.*, 37:117–126.
- Cuzzocrea, A., Martinelli, F., Mercaldo, F., and Vercelli, G. V. (2017). Tor traffic analysis and detection via machine learning techniques. In *IEEE BigData, 2017*, pages 4474–4480. IEEE Computer Society.
- Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *J. Big Data*, 6:54.
- Dimitsaki, S., Gavriilidis, G. I., Dimitriadis, V. K., and Natsiavas, P. (2023). Benchmarking of machine learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence. *Artif. Intell. Medicine*, 137:102490.
- Ding, J., Errapotu, S. M., Guo, Y., Zhang, H., Yuan, D., and Pan, M. (2022). Private empirical risk minimization with analytic gaussian mechanism for healthcare system. *IEEE Trans. Big Data*, 8(4):1107–1117.
- Elgammal, A. and Krämer, B. J. (2021). A reference architecture for smart digital platform for personalized prevention and patient management. In *Next-Gen Digital Services. A Retrospective and Roadmap for Service Computing of the Future - Essays Dedicated to Michael Papazoglou on the Occasion of His 65th Birthday and His Retirement*, volume 12521, pages 88–99.
- Fu, X., Zhang, B., Dong, Y., Chen, C., and Li, J. (2022). Federated graph machine learning: A survey of concepts, techniques, and applications. *SIGKDD Explor.*, 24(2):32–47.
- Ghayvat, H., Pandya, S., Bhattacharya, P., Zuhair, M., Rashid, M., Hakak, S., and Dev, K. (2022). CP-BDHCA: blockchain-based confidentiality-privacy preserving big data scheme for healthcare clouds and applications. *IEEE J. Biomed. Health Informatics*, 26(5):1937–1948.
- Gounaris, A. and Torres, J. (2018). A methodology for spark parameter tuning. *Big Data Res.*, 11:22–32.
- Leung, C. K., Cuzzocrea, A., Mai, J. J., Deng, D., and Jiang, F. (2019). Personalized deepinf: Enhanced social influence prediction with deep learning and transfer learning. In *IEEE BigData, 2019*, pages 2871–2880. IEEE.
- Lin, C., Song, Z., Song, H., Zhou, Y., Wang, Y., and Wu, G. (2016). Differential privacy preserving in big data analytics for connected health. *J. Medical Syst.*, 40(4):97:1–97:9.
- Liu, C., Yao, Z., Liu, P., Tu, Y., Chen, H., Cheng, H., Xie, L., and Xiao, K. (2023a). Early prediction of MODS interventions in the intensive care unit using machine learning. *J. Big Data*, 10(1):55.
- Liu, X., Hasan, M. R., Ahmed, K. A., and Hossain, M. Z. (2023b). Machine learning to analyse omic-data for COVID-19 diagnosis and prognosis. *BMC Bioinform.*, 24(1):7.
- Mohamed, M. A., El-Henawy, I. M., and Salah, A. (2021). Usages of spark framework with different machine learning algorithms. *Comput. Intell. Neurosci.*, 2021:1896953:1–1896953:7.
- Omran, N. F., Ghany, S. F. A., Saleh, H., and Nabil, A. (2021). Breast cancer identification from patients' tweet streaming using machine learning solution on spark. *Complex.*, 2021:6653508:1–6653508:12.
- Onesimu, J. A., Karthikeyan, J., Eunice, J., Pomplun, M., and Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access*, 10:86979–86997.
- Parimanam, K., Lakshmanan, L., and Palaniswamy, T. (2022). Hybrid optimization based learning technique for multi-disease analytics from healthcare big data using optimal pre-processing, clustering and classifier. *Concurr. Comput. Pract. Exp.*, 34(17).
- Patil, H. K. and Seshadri, R. (2014). Big data security and privacy issues in healthcare. In *IEEE Congress on Big Data, 2014*, pages 762–765. IEEE Computer Society.
- QUALITOP (2023). The QUALITOP project. <https://h2020qualitop.liris.cnrs.fr/wordpress/index.php/project/>.
- Singh, S., Rathore, S., Alfarraj, O., Tolba, A., and Yoon, B. (2022). A framework for privacy-preservation of iot healthcare data using federated learning and blockchain technology. *Future Gener. Comput. Syst.*, 129:380–388.
- Sun, J. and Reddy, C. K. (2013). Big data analytics for healthcare. In *ACM SIGKDD KDD, 2013*, page 1525. ACM.
- Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570.
- Tan, Q., Xu, X., and Liang, H. (2023). Physiological big data mining through machine learning and wireless sensor networks. *Int. J. Distributed Syst. Technol.*, 14(2):1–12.
- Vinke, P. C., Combalia, M., de Bock, G. H., Leyrat, C., Spanjaart, A. M., Dalle, S., da Silva, M. G., Essongue, A. F., Rabier, A., Pannard, M., Jalali, M. S., Elgammal, A., Papazoglou, M., Hacid, M.-S., Rioufol, C., Kersten, M.-J., van Oijen, M. G., Suazo-Zepeda, E., Malhotra, A., Coquery, E., Anota, A., Preau, M., Fauvernier, M., Coz, E., Puig, S., and Maucourt-Boulch, D. (2023). Monitoring multidimensional aspects of quality of life after cancer immunotherapy: protocol for the international multicentre, observational qualitop cohort study. *BMJ Open*, 13(4).
- Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):12:1–12:19.
- Zhang, H., Chen, G., Ooi, B. C., Tan, K., and Zhang, M. (2015). In-memory big data management and processing: A survey. *IEEE Trans. Knowl. Data Eng.*, 27(7):1920–1948.