Temporal Big Data Analytics: New Frontiers for Big Data Analytics Research

Alfredo Cuzzocrea[1]
iDEA Lab, University of Calabria, Rende, Italy
LORIA, Nancy, France

Abstract

Big data analytics is an emerging research area with many sophisticated contributions in the actual literature. Big data analytics aims at discovering actionable knowledge from large amounts of big data repositories, based on several approaches that integrate foundations of a wide spectrum of disciplines, ranging from data mining to machine learning and artificial intelligence. Among the concrete innovative topics of big data analytics, *temporal big data analytics* covers a first-class role and it is attracting the attention of larger and larger communities of academic and industrial researchers. Basically, temporal big data analytics aims at modeling, capturing and analyzing temporal aspects of big data during analytics phase, including specialized tasks such as *big data versioning over time*, building temporal relations among ad-hoc *big data structures* (such as nodes of *big graphs*) and *temporal queries over big data*. It is worth to notice that temporal big data analytics research is characterized by several open challenges, which range from foundations, including temporal big data representation and processing, to applications, including smart cities and bio-informatics tools. Inspired by these considerations, this paper focuses on models, paradigms, techniques and future challenges of temporal big data analytics, by reporting on state-of-the-art results as well as emerging trends, with also criticisms on future work that we should expect from the community.

2012 ACM Subject Classification Information systems → Temporal data; Information systems → Data analytics

Keywords and phrases Big Data Analytics, Big Data Management, Temporal Big Data Analytics, Temporal Big Data Management

Digital Object Identifier 10.4230/LIPIcs.TIME.2021.4

Category Panel Description

Funding This research has been partially supported by the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

Acknowledgements The open access publication of this article was supported by the Alpen-Adria-Universität Klagenfurt, Austria.

1 Introduction

Big data analytics is an emerging research area with many sophisticated contributions in the actual literature (e.g., [28, 18, 7, 21]). Big data analytics aims at discovering actionable knowledge from large amounts of big data repositories, based on several approaches that integrate foundations of a wide spectrum of disciplines, ranging from data mining to machine learning and artificial intelligence. Among the concrete innovative topics of big data analytics, *temporal big data analytics* covers a first-class role and it is attracting the attention of larger and larger communities of academic and industrial researchers (e.g., [9, 26, 10, 24, 32]). Basically, temporal big data analytics aims at modeling, capturing and analyzing temporal

[1] This research has been made in the context of the Excellence Chair in Computer Engineering – Big Data Management and Analytics at LORIA, Nancy, France

© Alfredo Cuzzocrea;
licensed under Creative Commons License CC-BY 4.0

28th International Symposium on Temporal Representation and Reasoning (TIME 2021).
Editors: Carlo Combi, Johann Eder, and Mark Reynolds; Article No. 4; pp. 4:1–4:7
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

aspects of big data during analytics phase, including specialized tasks such as *big data versioning over time*, building temporal relations among ad-hoc *big data structures* (such as nodes of *big graphs*) and *temporal queries over big data.* It is worth to notice that temporal big data analytics research is characterized by several open challenges, which range from foundations, including temporal big data representation and processing, to applications, including smart cities and bio-informatics tools.

Temporal big data analytics arise in many application scenarios. Consider, for instance, the case of social networks. Here, the user content (such as posts, pictures, videos, etc.) identify a very large big data repository, while the relationships among users and contents, users and users, etc., across time, define an *evolving* temporal big data set. Basically, for each reference timestamp, namely $t_i, t_{i+1}, t_{i+2}, \ldots$, we have a different big data set, namely $\mathcal{B} = \{B(t_i), B(t_{i+1}), B(t_{i+2}), \ldots\}$. How to make analytics *across* the various big data sets $B(t_i), B(t_{i+1}), B(t_{i+2}), \ldots$ in $\mathcal{B}$? For instance, what was the most frequent itemset pattern of posts of the user Bob at timestamp t_j and what the one at timestamp t_{j+k}? Furthermore, what is the intersection set of Facebook common friends of Bob and Alice at timestamp t_j and what the one at timestamp t_{j+k}? How common friends ave evolved (e.g., removing old friends or adding new friends) across time? The latter one are simplest temporal big data analytics patterns, which can be further developed towards more complex ones, for instance based on *multidimensional analytics* (e.g., [19, 8, 11]).

In all the described settings, *temporal big queries* play a critical role. How to query "historical" big data repositories? In order to to this, first a suitable representation model for capturing the temporal aspect of big data must be introduced. The most popular approach to this end is presented by straightforward application of classical models for temporal databases, but targeted to the specific big data environment, thus, considering, for instance, *scalability issues* (e.g., [22, 2, 20]). Here, big data objects are associated with a proper timestamp, according to three different schema, namely *valid time*, *transaction time*, and *decision time.* According to the first schema, given a big data object O_k, the reference timestamp associated to O_k, namely t_k, is the time period during which a fact is true in the real world. According to the second schema, t_k is the time at which a fact was recorded in the database. Finally, according to the third schema, t_k is the time at which the decision was made about the fact. Based on the given temporal data model, given a big data repository B, a temporal big query over B referring to timestamp t_k, denoted by $Q(B)_{t_k}$, retrieves from B all the big data objects in B associated to the timestamp t_k and that satisfy the query predicates in Q (e.g., aggregation predicates – [30, 15]). Formally, a temporal big data analytics $\mathcal{A}$ over B, is defined as a collection of analytical functions $\mathcal{A} = \{f_0, f_1, f_2, \ldots, f_{n-1}\}$ such that every analytical function f_j is defined on top of a collection of a set of temporal big queries $\mathcal{Q} = \{Q(B)_{t_k}, Q(B)_{t_{k+1}}, Q(B)_{t_{k+2}}, \ldots Q(B)_{t_{k+m-1}}\}$. f_j can be of different nature, for instance based on *OLAP analytics* (e.g., [6, 16, 13]).

The above described model based on temporal big queries can be easily extended to more complex procedures, including, for instance, *data mining programs* and *machine learning procedures*, still parameterized by timestamp. In this vest, the integration of the latter temporal big data analytics model with emerging *NoSQL databases* (e.g., [29]) turns to be very interesting.

Inspired by these considerations, this paper focuses on models, paradigms, techniques and future challenges of temporal big data analytics, by reporting on state-of-the-art results as well as emerging trends, with also criticisms on future work that we should expect from the community. The remaining part of this paper is organized as follows. In Section 2, we provide a brief overview of most relevant state-of-the-art proposal for temporal big data analytics appearing in literature. Section 3 is devoted to the description of most relevant emerging challenges and open issues for temporal big data analytics. Finally, in Section 4, we derive conclusions of our research.

2 Temporal Big Data Analytics: State-Of-The-Art Proposals

Inspired by considerations reported above, temporal big data analytics is a rich area of research. As a consequence, there exist in literature a number of relevant proposals, with many interesting scientific as well as industrial outcomes. Here, we outline some of the most relevant ones.

[9] focuses on the issue of supporting *temporal analytics on big data for web advertising*. For instance, they consider display advertising that makes use of Behavioral Targeting (BT) to select ads for users based on prior searches, page views, etc. Previous work on BT has focused on techniques that scale well for offline data using Map-Reduce (M-R). However, this approach has limitations for BT-style applications that deal with temporal data: (1) many queries are temporal and not easily expressible in M-R, and moreover, the set-oriented nature of M-R frontends such as SCOPE is not suitable for temporal processing; (2) as commercial systems mature, they may need to also directly analyze and react to real-time data feeds since a high turnaround time can result in missed opportunities, but it is difficult for current solutions to naturally also operate over real-time streams. The contributions of the paper are twofold. First, authors propose a novel framework called *TiMR*, that combines a time-oriented data processing system with a M-R framework. Users perform analytics using temporal queries – these queries are succinct, scale-out-agnostic, and easy to write. They scale well on large-scale offline data using TiMR, and can work unmodified over real-time streams. They also propose new cost-based query fragmentation and temporal partitioning schemes for improving efficiency with TiMR. Second, they show the feasibility of this approach for BT, with new temporal algorithms that exploit new targeting opportunities. Experiments using real advertising data show that TiMR is efficient and incurs orders-of-magnitude lower development effort. The proposed BT solution is easy and succinct, and performs up to several times better than current schemes in terms of memory, learning time, and click-through-rate/coverage.

[26] considers, instead, *temporal event tracing on big healthcare data analytics*. The study presents a comprehensive method for rapidly processing, storing, retrieving, and analyzing big healthcare data. Based on NoSQL, a patient-driven data architecture is suggested to enable the rapid storing and flexible expansion of data. Thus, the schema differences of various hospitals can be overcome, and the flexibility for field alterations and addition is ensured. The timeline mode can easily be used to generate a visual representation of patient records, providing physicians with a reference for patient consultation. The sharding-key is used for data partitioning to generate data on patients of various populations. Subsequently, data reformulation is conducted as a first step, producing additional temporal and spatial data, providing cloud computing methods based on query-MapReduce-shard, and enhancing the search performance of data mining. Target data can be rapidly searched and filtered, particularly when analyzing temporal events and interactive effects.

[10] deals with the challenge of defining and implementing *temporal data analytics on COVID-19 data with ubiquitous computing*. Authors argue that, with technological advancements in computing and communications, huge amounts of big data are generated and collected at a very rapid rate from a wide variety of rich data sources. Embedded in these big data are useful information and valuable knowledge. An example is healthcare and

epidemiological data such as data related to patients who suffered from viral diseases like the coronavirus disease 2019 (COVID-19). Knowledge discovered from these epidemiological data via data science helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. In the paper, authors present a temporal data science algorithm for analyzing big COVID-19 epidemiological data, with focus on the temporal data analytics with ubiquitous computing. The algorithm helps users to get a better understanding of information about the confirmed cases of COVID-19. Evaluation results show the benefits of the proposed system in temporal data analytics of big COVID-19 data with ubiquitous computing. Although the algorithm is designed for temporal data analytics of big epidemiological data, it would be applicable to other temporal data analytics of big data in many real-life applications and services.

[24] moves the attention on *large-scale smart grids, with specific temporal, functional and spatial big data computing features*. Indeed, with the deployment of monitoring devices, the smart grid is collecting large amounts of energy-related data at an unprecedented speed. The smart grid has become data-driven, which necessitates extracting meaningful data from a large dataset. The traditional approach of data extraction improves the computing efficiency in temporal dimension, but it is made for only one task in the smart grid. Moreover, the existing solutions neglect the geographical distribution of computing capacity in a large-scale smart grid. The future large-scale smart grid will run over the internet of energy where the dataset will be sent to a specific destination along power routers hop-by-hop. Consequently, authors design a novel temporal, functional and spatial big data computing framework for large-scale smart grid. In functional dimension, they divide every dataset into sub-groups, each of which has data items shared by different tasks. In spatial dimension, they determine which location the power router should be placed to harvest computing resources used for extracting the sub-group of data items. The proposed method achieves a promising computing efficiency approaching to the optimal solution with 95 percent convergence ratio, and it saves the in-path bandwidth with 81 percent improvement ratio over benchmarks.

Finally, [32] considers the specific application scenario represented by *scientific big data analytics* and, in particular, *text-based temporally linked event networks* in such a scenario. Authors recognize that events formulate the world of the human being and could be regarded as the semantic units in different granularities for information organization. Extracting events and temporal information from texts plays an important role for information analytics in big data because of the wide use of multilingual texts. Based on these considerations, the paper surveys existing research work on text-based event temporal resolution and reasoning including identification of events, temporal information resolutions of events in English and Chinese texts, the rule-based temporal relation reasoning between events and relevant temporal representations. For the scientific big data analytics, authors point out the shortcomings of existing research work and give the argument about the future research work for advancing identification of events, establishment of temporal relations and reasoning of temporal relations.

3 Temporal Big Data Analytics: Emerging Challenges and Open Issues

Temporal big data analytics opens several new and challenging research perspectives. In the next, we focus the attention on those that have been scored as relevant by our study.

Big Temporal Data Representation. As mentioned in Section 1, the first issue to face-off is represented by how to model and capture big temporal data for supporting temporal big data analytics effectively and efficiently. This involves in devising suitable *temporal big data models*, which must also consider scalability issue of big temporal data processing (e.g., [5]). On the other hand, another relevant aspect of big temporal data to consider is their clear *multi-granularity nature*, according to which the same (big) data appear in different scales and resolutions (e.g., year, quarter, month, and so forth – [4]). How to represent this special feature effectively and efficiently? Cloud computing environments, with well-known *elastic metaphors* (e.g., [1]), seem to be the most promising computational solutions to be considered by future research efforts.

Big Queries on Big Temporal Data Repositories. Big queries on big temporal data repositories is an exciting new challenge that is critical for temporal big data analytics, like in some related cases (e.g., [31]). Basically, considering that big temporal data repositories are characterized by a *strong heterogeneity*, big queries appear in several formats (e.g., tree-like queries, graph-like queries, and so forth) and, as a consequence, the big query optimization and evaluation layer of a hypothetical temporal big data analytics engine should include several characteristics and functionalities, such as query translation and query rewriting (e.g., [23]). In this respect, the implementation on top of MapReduce framework seems a promising direction to follow.

Machine-Learning-Based Temporal Big Data Analytics. Temporal big data analytics are usually based on a *core methodology* that characterizes their analytical functions (see Section 1). Among several alternatives, *machine-learning-based temporal big data analytics* should be considered as the most sophisticated collection of techniques available in literature (e.g., [27]). Indeed, machine learning techniques are flexible enough to deal with big temporal data, and to support the relevant knowledge discover process from such big data repositories effectively and efficiently. A wide family of proposals are available in actual literature, and they can be applied to the issue of supporting temporal big data analytics in real-life application scenarios (e.g., smart cities, sensor networks, IoT systems, and so forth).

Uncertainty and Imprecision in Temporal Big Data Analytics. Big temporal data are naturally affected by *uncertainty and imprecision* (e.g., [17]), due to several reasons, among which data transmission errors and human data entry errors are just two possible instances. How uncertainty and imprecision of big temporal data impact on the accuracy and the global-quality of temporal big data analytics procedures built on top of them? This is a critical question that future research efforts must consider seriously. To this end, *probabilistic temporal big data analytics models* (e.g., [12]) seem to be a solid paradigm to be considered.

Privacy-Preserving Temporal Big Data Analytics. Temporal big data analytics usually access *sensitive data*. To be convinced of this, consider the case of temporal big data analytics developed in the contest of federated healthcare systems. Here, a massive amount of *personal data* (e.g., [3]) are accessed and processed in order to derive suitable analytics for decision making. How to design and devise *privacy-preserving temporal big data analytics* models capable of preserving the privacy of sensitive (e.g., personal) data while not lowering the degree of accuracy and decision-making-support of the target temporal big data analytics? The latter question will be an annoying challenge for many years to come (e.g., [14]).

4 Conclusions

In this paper, we provided a comprehensive overview of state-of-the-art temporal big data analytics techniques and algorithms, by highlighting their benefits and limitations. As a further contribution of our work, we have also provided a discussion on open research challenges and future directions in this scientific field, aiming at achieving a significant milestone to be exploited by forthcoming research efforts. Last but not least, we firmly believe that innovative and emerging application scenarios (e.g., [25]) will provide more insights and inspiration to the research community for the future years.

References

1 Divyakant Agrawal, Amr El Abbadi, Sudipto Das, and Aaron J Elmore. Database scalability, elasticity, and autonomy in the cloud. In *International Conference on Database Systems for Advanced Applications*, pages 2–15. Springer, 2011.

2 Sara Alghunaim and Heyam H Al-Baity. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access*, 7:91535–91546, 2019.

3 Rasim M Alguliyev, Ramiz M Aliguliyev, and Fargana J Abdullayeva. Privacy-preserving deep learning algorithm for big personal data analysis. *Journal of Industrial Information Integration*, 15:1–14, 2019.

4 Safaa Alwajidi and Li Yang. Multi-resolution hierarchical structure for efficient data aggregation and mining of big data. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pages 153–159. IEEE, 2019.

5 Ladjel Bellatreche, Alfredo Cuzzocrea, and Soumia Benkrid. $\mathcal{F}\&\mathcal{A}$: A methodology for effectively and efficiently designing parallel relational data warehouses on heterogenous database clusters. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 89–104. Springer, 2010.

6 Boualem Benatallah, Hamid Reza Motahari-Nezhad, et al. Scalable graph-based olap analytics over process execution data. *Distributed and Parallel Databases*, 34(3):379–423, 2016.

7 Alina Campan, Alfredo Cuzzocrea, and Traian Marius Truta. Fighting fake news spread in online social networks: Actual trends and future research directions. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4453–4457. IEEE, 2017.

8 Michelangelo Ceci, Alfredo Cuzzocrea, and Donato Malerba. Effectively and efficiently supporting roll-up and drill-down olap operations over continuous dimensions via hierarchical clustering. *Journal of Intelligent Information Systems*, 44(3):309–333, 2015.

9 Badrish Chandramouli, Jonathan Goldstein, and Songyun Duan. Temporal analytics on big data for web advertising. In *2012 IEEE 28th international conference on data engineering*, pages 90–101. IEEE, 2012.

10 Yubo Chen, Carson K Leung, Siyuan Shang, and Qi Wen. Temporal data analytics on covid-19 data with ubiquitous computing. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 958–965. IEEE, 2020.

11 Alfredo Cuzzocrea. Accuracy control in compressed multidimensional data cubes for quality of answer-based olap tools. In *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*, pages 301–310. IEEE, 2006.

12 Alfredo Cuzzocrea. Retrieving accurate estimates to olap queries over uncertain and imprecise multidimensional data streams. In *International Conference on Scientific and Statistical Database Management*, pages 575–576. Springer, 2011.

13 Alfredo Cuzzocrea. Analytics over big data: Exploring the convergence of datawarehousing, olap and data-intensive cloud infrastructures. In *2013 IEEE 37th Annual Computer Software and Applications Conference*, pages 481–483. IEEE, 2013.

14 Alfredo Cuzzocrea. Privacy and security of big data: current challenges and future research perspectives. In *Proceedings of the first international workshop on privacy and secuirty of big data*, pages 45–47, 2014.

15 Alfredo Cuzzocrea. Aggregation and multidimensional analysis of big data for large-scale scientific applications: models, issues, analytics, and beyond. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2015.

16 Alfredo Cuzzocrea, Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Mimmo Parente. Olap analysis of multidimensional tweet streams for supporting advanced analytics. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 992–999, 2016.

17 Alfredo Cuzzocrea, Carson Kai-Sang Leung, and Richard Kyle MacKinnon. Mining constrained frequent itemsets from distributed uncertain data. *Future Generation Computer Systems*, 37:117–126, 2014.

18 Alfredo Cuzzocrea and Il-Yeol Song. Big graph analytics: the state of the art and future research agenda. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP*, pages 99–101, 2014.

19 Alfredo Cuzzocrea, Il-Yeol Song, and Karen C Davis. Analytics over large-scale multidimensional data: the big data revolution! In *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, pages 101–104, 2011.

20 Alfredo Cuzzocrea and Wei Wang. Approximate range–sum query answering on data cubes with probabilistic guarantees. *Journal of Intelligent Information Systems*, 28(2):161–197, 2007.

21 Timothée Dubuc, Frederic Stahl, and Etienne B Roesch. Mapping the big data landscape: technologies, platforms and paradigms for real-time analytics of data streams. *IEEE Access*, 9:15351–15374, 2020.

22 Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, and Francisco Herrera. A comparison on scalability for batch big data processing on apache spark and apache flink. *Big Data Analytics*, 2(1):1–11, 2017.

23 Rihan Hai, Christoph Quix, and Chen Zhou. Query rewriting for heterogeneous data lakes. In *European Conference on Advances in Databases and Information Systems*, pages 35–49. Springer, 2018.

24 Weigang Hou, Zhaolong Ning, Lei Guo, and Xu Zhang. Temporal, functional and spatial big data computing framework for large-scale smart grid. *IEEE Transactions on Emerging Topics in Computing*, 7(3):369–379, 2017.

25 Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung. Big-data applications in the government sector. *Communications of the ACM*, 57(3):78–85, 2014.

26 Chin-Ho Lin, Liang-Cheng Huang, Seng-Cho T Chou, Chih-Ho Liu, Han-Fang Cheng, and I-Jen Chiang. Temporal event tracing on big healthcare data analytics. In *2014 IEEE International Congress on Big Data*, pages 281–287. IEEE, 2014.

27 Ives Cavalcante Passos, Benson Mwangi, and Flávio Kapczinski. Big data analytics and machine learning: 2015 and beyond. *The Lancet Psychiatry*, 3(1):13–15, 2016.

28 Philip Russom et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.

29 Michael Stonebraker. Sql databases v. nosql databases. *Communications of the ACM*, 53(4):10–11, 2010.

30 Allen Van Gelder. The well-founded semantics of aggregation. In *Proceedings of the Eleventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 127–138, 1992.

31 Xiaochun Yun, Guangjun Wu, Guangyan Zhang, Keqin Li, and Shupeng Wang. Fastraq: A fast approach to range-aggregate queries in big data environments. *IEEE Transactions on Cloud Computing*, 3(2):206–218, 2014.

32 Junsheng Zhang, Changqing Yao, Yunchuan Sun, and Zengquan Fang. Building text-based temporally linked event network for scientific big data analytics. *Personal and Ubiquitous Computing*, 20(5):743–755, 2016.