Telco Big Data: Current State & Future Directions

Constantinos Costa and Demetrios Zeinalipour-Yazti
Department of Computer Science
University of Cyprus, 1678 Nicosia, Cyprus
{costa.c, dzeina}@cs.ucy.ac.cy

***Abstract*—A Telecommunication company (Telco) is traditionally only perceived as the entity that provides telecommunication services, such as telephony and data communication access to users. However, the radio and backbone infrastructure of such entities spanning densely most urban spaces and widely most rural areas, provides nowadays a unique opportunity to collect immense amounts of data that capture a variety of natural phenomena on an ongoing basis, e.g., traffic, commerce and mobility patterns and user service experience. The ability to perform analytics on the generated big data within tolerable elapsed time and share it with key smart city enablers (e.g., municipalities, public services, startups, authorities, and companies), elevates the role of Telcos in the realm of future smart cities from pure network access providers to information providers. In this talk, we overview the state-of-the-art in Telco big data analytics by focusing on a set of basic principles, namely: (i) real-time analytics and detection; (ii) experience, behavior and retention analytics; (iii) privacy; and (iv) storage. We also present experiences from developing an innovative such architecture and conclude with open problems and future directions.**

***Keywords*—Telco, Telecommunication, Big Data, Queries, Analytics, Storage, Privacy**

I. Introduction

Unprecedented amounts and variety of spatiotemporal *big* data are generated every few minutes by the infrastructure of a *telecommunication company (telco)*. The rapid expansion of broadband mobile networks, the pervasiveness of smartphones, and the introduction of dedicated Narrow Band connections for smart devices and Internet of Things (NB-IoT) [1] have contributed to this explosion. For example, a telco in the city of Shenzhen, China, which serves 10 million users produce 5TB per day [2] (i.e., thousands to millions of records every second). Huang et al. [3] break their 2.26TB per day *Telco Big Data (TBD)* down as follows: (i) *Business Supporting Systems (BSS)* data, which is generated by the internal work-flows of a telco (e.g., billing, support), accounting to a moderate of 24GB per day and; (ii) *Operation Supporting Systems (OSS)* data, which is generated by the Radio and Core equipment of a telco, accounting to 2.2TB per day and occupying over 97% of the total volume.

Data exploration queries over big telco data are of great interest to both the telco operators and the smart city enablers (e.g., municipalities, public services, startups, authorities, and companies), as these allow for interactive analysis at various granularities, narrowing it down for a variety of tasks. Effectively storing and processing TBD workflows can unlock a wide spectrum of challenges, ranging from churn prediction of subscribers [3], city localization [4], 5G network optimization / user-experience assessment [5]–[7] and road traffic mapping [8]. Data exploration and visualization might be the most important tools in the big data era [9]–[11], where decision support makers, ranging from CEOs to front-line support engineers, aim to draw valuable insights and conclusions visually.

Our tutorial will tackle the topic from a wide range of perspectives: fundamentals, definitions, current state, academic & industrial perspective, reality & visionary scenarios as well as future challenges. The seminar captures the big picture, such that interested researchers and practitioners can expand their study by following the references. Our presentation is carried out through the lens of an experimental Telco Big Data System we developed at the University of Cyprus, coined SPATE [6], which is a SPAtio-TEmporal framework that uses both lossless data *compression* and lossy data *decaying* (i.e., *Data Postdiction* [12]) to ingest large quantities of telco big data in the most compact manner.

Compression refers to the encoding of data using fewer bits than the original representation and is important as it shifts the resource bottlenecks from storage- and network-I/O to CPU, whose cycles are increasing at a much faster pace [13]–[15]. It also enables data exploration tasks to retain full resolution over the most important collected data. *Decaying* on the other hand, as suggested in [16], refers to the progressive loss of detail in information as data ages with time until it has completely disappeared (the schema of the database does not decay [17]). This enables data exploration tasks to retain high-level data exploration capabilities for predefined aggregates over long time windows, without consuming enormous amounts of storage.

Our tutorial aims to provide an extensive coverage of telco big data research, which falls under the following categories: (i) real-time analytics and detection; (ii) experience, behavior and retention analytics; (iii) privacy; and (iv) storage. There is also traditional telco research not related to big data, rather comprises of topics related to business (BSS) data in relational databases. The given presentation should allow the audience to grasp basic and advanced concepts ranging from the anatomy of a telco network and the structure of telco big data all the way up to applications and benefits of Telco Big Data. We will conclude the seminar with the presentation of the challenges and opportunities in the field.

To our knowledge, this is the first tutorial covering explicitly telco big data and this stems directly from our recent work on the subject covered in [6] [18] [12] and the Telco Big Data Awareness project[1]. The intended duration of our tutorial is 1.5 hours and we look forward to between 25-50 attendees at the conference. This is the first time this tutorial will be presented at a conference and we believe it will create vibrant discussions with attendees at the conference.

[1]TBD Awareness. https://tbd.cs.ucy.ac.cy/

OUTLINE

In this section we outline the tentative structure of the advanced seminar during the conference. The final layout of the seminar will be reflected in its power-point presentation available through the seminar website [2].

Real-time Analytics and Detection: Zhang et al. [2] developed *OceanRT*, which was one of the first real-time telco big data analytic demonstrations. Yuan et al. [19] present *OceanST* which features: (i) an efficient loading mechanism of ever-growing telco MBB data; (ii) new spatiotemporal index structures to process exact and approximate spatiotemporal aggregate queries. Iyer et al. [5] present *CellIQ* to optimize queries such as *"spatiotemporal traffic hotspots"* and *"handoff sequences with performance problems"*. It represents the snapshots of cellular network data as graphs and leverages on the spatial and temporal locality of cellular network data. Zhu et al. [4] deal with the usage of telco MR data for city-scale localization, which is complementary to the scope of our work.

Braun et al. [20] develop a scalable distributed system that efficiently processes mixed workloads to answer event stream and analytic queries over telco data. Bouillet et al. [21] develop a system on top of IBM's InfoSphere Streams middleware that analyzes 6 billion CDR per day in real-time. Abbasoğlu et al. [22] present a system for maintaining call profiles of customers in a streaming setting by applying scalable distributed stream processing.

Experience, Behavior and Retention Analytics: Huang et al. [3] empirically demonstrate that customer churn prediction performance can be significantly improved with telco big data. Although BSS data have been utilized in churn prediction very well in the past decade, the authors show how with a primitive Random Forest classifier telco big data can improve churn prediction accuracy from 68% to 95%. Luo et al. [7] propose a framework to predict user behavior involving more than one million telco users. They represent users as documents containing a collection of changing spatiotemporal "words" that express user behavior. By extracting the users' space-time access records from MBB data, they learn user-specific compact topic features that they use for user activity level prediction. Ho et. al. [23] propose a distributed community detection algorithm that aims to discover groups of users that share similar edge properties reflecting customer behavior.

Privacy: Hu et al. [24] study Differential Privacy for data mining applications over telco big data and show that for real-word industrial data mining systems the strong privacy guarantees given by differential privacy are traded with a 15% to 30% loss of accuracy. Privacy and confidentiality are critical for telcos' reliability due to the highly sensitive attributes of user data located in CDR, such as billing records, calling numbers, call duration, data sessions, and trajectory information. *SPATE* deals with privacy-aware data sharing as a functionality for next generation smart-city applications.

Storage: One key challenge in this new era of telco big data is to *minimize the storage costs* associated with the data exploration tasks, as big data traces and computed indexes can have a tremendous storage and I/O footprint on the data centers of telcos. Storing big data locally, due to the sensitive nature of data that cannot reside on public cloud storage, adds great challenges and costs that reach beyond the simplistic capacity cost calculated per GB [25]. From a telco's perspective, the requirement is to: (i) *incrementally store big data in the most compact manner*, and (ii) *improve the response time for data exploration queries*. These two objectives are naturally conflicting, as conjectured in [26]. In previous work, custom data management systems have been designed with the objectives to save storage space using compression, and speed up temporal range queries using indices [27]–[30]. None of these considers the notion of "decay" as expressed in [16], which suggests sacrificing either accuracy or read efficiency for less frequently accessed data to save space.

[2]Seminar slides: https://dmsl.cs.ucy.ac.cy/tutorials/mdm18/

II. DESCRIPTION OF TARGET AUDIENCE

The goal of this advanced seminar is to convey a basic and advanced understanding of the unique characteristics, challenges and opportunities of telco big data management and how these can facilitate Mobile Data Management research, evaluation and applications. The advanced seminar is targeted to scientists with a basic understanding of mobile data management, but no knowledge of telco data management technologies is required. In particular, this seminar addresses the following audience:

- Graduate and Undergraduate Students
- Mobile Data Management Researchers/Educators
- Industry Developers

This seminar covers, but is not limited to, the following MDM 2018 topics of interest:

- Data Management for Internet of Things (IoT) and Sensor Systems
- Data Management for Connected Cars, Intelligent Transportation Systems, Smart Spaces
- Mobile Crowd-Sourcing and Crowd-Sensing
- Mobile Data Analytics
- Behavioral/Activity Sensing and Analytics
- Middleware and Tools for Mobile and Pervasive Computing
- Data Stream Processing in Mobile/Sensor Network
- Indexing, Optimisation and Query Processing for Moving Objects/Users
- Location and Trajectory Analytics
- Routing, Personalized Routing, Eco-Routing, Routing for Electrical Vehicles
- Innovative Applications driven by Mobile Data

BIOGRAPHIES OF ADVANCED SEMINAR PRESENTERS

Constantinos Costa is a full-time Ph.D. Candidate and a Research Assistant at the Department of Computer Science (UCY), being involved in research at the Data Management Systems Laboratory (DMSL). He holds a M.Sc. degree in Computer Science (2013) and a B.Sc. degree in Computer Science (2011) from the University of Cyprus. His research interests include databases and mobile computing, particularly distributed query processing for spatial and spatio-temporal datasets. Costa has contributed extensively to open source projects for indoor navigation, crowd messaging and telco big data. For more information please visit: https://www.cs.ucy.ac.cy/~costa.c/.

Demetrios Zeinalipour-Yazti is an Associate Professor of Computer Science at the University of Cyprus, where he leads the Data Management Systems Laboratory (DMSL). His primary research interests include Data Management in Computer Systems and Networks, particularly Mobile and Sensor Data Management; Big Data Management in Parallel and Distributed Architectures; Spatio-Temporal Data Management; Network and Telco Data Management; Crowd, Web 2.0 and Indoor Data Management; Data Privacy Management. He holds a Ph.D. in Computer Science from University of California - Riverside (2005). Before his current appointment, he served the University of Cyprus as an Assistant Professor and Lecturer but also the Open University of Cyprus as a Lecturer. He has held visiting research appointments at Akamai Technologies, Cambridge, MA, USA, the University of Athens, Greece, the University of Pittsburgh, PA, USA and the Max Planck Institute for Informatics, Saarbrücken, Germany. He is a Humboldt Fellow, Marie-Curie Fellow, an ACM Distinguished Speaker (2017-2020), a Senior Member of ACM, a Senior Member of IEEE and a Member of USENIX. He serves on the editorial board of Distributed and Parallel Databases (Elsevier), Big Data Research (Springer) and is an independent evaluator for the European Commission (Marie Skodowska-Curie and COST actions).

His h-index is 24, holds over 2600 citations, has an Erds number of 3, won 10 international awards (ACMD17, ACMS16, IEEES16, HUMBOLDT16, IPSN14, EVARILOS14, APPCAMPUS13, MDM12, MC07, CIC06) and delivered over 30 invited talks. He has participated in over 20 projects funded by the US National Science Foundation, by the European Commission, the Cyprus Research Promotion Foundation, the Univ. of Cyprus, the Open University of Cyprus and the Alexander von Humboldt Foundation, Germany. Finally, he has also been involved in industrial Research and Development projects (e.g., Finland, Taiwan and Cyprus) and has technically lead several mobile data management services (e.g., Anyplace, Rayzit and Smartlab) reaching over 35K users worldwide with over 140K sessions. For more information please visit: https://www.cs.ucy.ac.cy/~dzeina/ or the DMSL website: https://dmsl.cs.ucy.ac.cy/.

REFERENCES

[1] Ericsson.com, "Cellular Networks For Massive IoT enabling low power wide area applications," 2016. [Online]. Available: https://goo.gl/Sf2Cj4

[2] S. Zhang, Y. Yang, W. Fan, L. Lan, and M. Yuan, "Oceanrt: Real-time analytics over large temporal data," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 1099–1102.

[3] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang, and J. Zeng, "Telco churn prediction with big data," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD. New York, NY, USA: ACM, 2015, pp. 607–618.

[4] F. Zhu, C. Luo, M. Yuan, Y. Zhu, Z. Zhang, T. Gu, K. Deng, W. Rao, and J. Zeng, "City-scale localization with telco big data," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM. New York, NY, USA: ACM, 2016, pp. 439–448.

[5] A. P. Iyer, L. E. Li, and I. Stoica, "Celliq: Real-time cellular network analytics at scale," in *Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'15. Berkeley, CA, USA: USENIX Association, 2015, pp. 309–322.

[6] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Efficient exploration of telco big data with compression and decaying," in *33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017*, 2017, pp. 1332–1343. [Online]. Available: https://doi.org/10.1109/ICDE.2017.175

[7] C. Luo, J. Zeng, M. Yuan, W. Dai, and Q. Yang, "Telco user activity level prediction with massive mobile broadband data," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 63:1–63:30, May 2016.

[8] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Towards real-time road traffic analytics using telco big data," in *Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, BIRTE, Munich, Germany, August 28, 2017*, 2017, pp. 5:1–5:5. [Online]. Available: http://doi.acm.org/10.1145/3129292.3129296

[9] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[10] TeraLab, "TeraLab Data Science for Europe," 2016. [Online]. Available: http://www.teralab-datascience.fr/

[11] A. Eldawy, M. F. Mokbel, S. Alharthi, A. Alzaidy, K. Tarek, and S. Ghani, "Shahed: A mapreduce-based system for querying and visualizing spatio-temporal satellite data," in *2015 IEEE 31st International Conference on Data Engineering*, April 2015, pp. 1585–1596.

[12] C. Costa, A. Charalampous, A. Konstantinidis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Decaying telco big data with data postdiction," in *Proceedings of the 19th IEEE International Conference on Mobile Data Management, June 25 - June 28, 2018, AAU, Aalborg, Denmark (accepted)*, 2018, p. 10 pages.

[13] Y. Chen, A. Ganapathi, and R. H. Katz, "To compress or not to compress - compute vs. io tradeoffs for mapreduce energy efficiency," in *Proceedings of the First ACM SIGCOMM Workshop on Green Networking*, ser. Green Networking '10, 2010, pp. 23–28.

[14] B. Welton, D. Kimpe, J. Cope, C. M. Patrick, K. Iskra, and R. Ross, "Improving i/o forwarding throughput with data compression," in *2011 IEEE Intl. Conference on Cluster Computing*, Sept 2011, pp. 438–445.

[15] T. Bicer, J. Yin, D. Chiu, G. Agrawal, and K. Schuchardt, "Integrating online compression to accelerate large-scale data analytics applications," in *Parallel Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, May 2013, pp. 1205–1216.

[16] M. L. Kersten, "Big data space fungus," in *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[17] M. Stonebraker, R. Castro, F. Dong Deng, and M. Brodie, "Database decay and what to do about it." 2016. [Online]. Available: https://goo.gl/tJNa9m

[18] D. Chatziantoniou, M. Castellanos, and P. K. Chrysanthis, Eds., *Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, BIRTE, Munich, Germany, August 28, 2017*. ACM, 2017. [Online]. Available: http://doi.acm.org/10.1145/3129292

[19] M. Yuan, K. Deng, J. Zeng, Y. Li, B. Ni, X. He, F. Wang, W. Dai, and Q. Yang, "Oceanst: A distributed analytic system for large-scale spatiotemporal mobile broadband data," *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1561–1564, Aug. 2014.

[20] L. Braun, T. Etter, G. Gasparis, M. Kaufmann, D. Kossmann, D. Widmer, A. Avitzur, A. Iliopoulos, E. Levy, and N. Liang, "Analytics in motion: High performance event-processing and real-time analytics in the same database," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD. New York, NY, USA: ACM, 2015, pp. 251–264.

[21] E. Bouillet, R. Kothari, V. Kumar, L. Mignet, S. Nathan, A. Ranganathan, D. S. Turaga, O. Udrea, and O. Verscheure, "Processing 6 billion cdrs/day: From research to production (experience report)," in *Proceedings of the 6th ACM Intl. Conference on Distributed Event-Based Systems*, ser. DEBS, 2012, pp. 264–267.

[22] M. A. Abbasoğlu, B. Gedik, and H. Ferhatosmanoğlu, "Aggregate profile clustering for telco analytics," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1234–1237, Aug. 2013.

[23] Q. Ho, W. Lin, E. Shaham, S. Krishnaswamy, T. A. Dang, J. Wang, I. C. Zhongyan, and A. She-Nash, "A distributed graph algorithm for discovering unique behavioral groups from large-scale telco data," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM. New York, NY, USA: ACM, 2016, pp. 1353–1362.

[24] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, and J. Zeng, "Differential privacy in telco big data platform," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1692–1703, Aug. 2015.

[25] Z. Li, A. Mukker, and E. Zadok, "On the importance of evaluating storage systems' $costs," in *6th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage)*, Philadelphia, PA, 2014.

[26] M. Athanassoulis, M. S. Kester, L. M. Maas, R. Stoica, S. Idreos, A. Ailamaki, and M. Callaghan, "Designing access methods: The rum conjecture," in *Intl. Conf. on Ext. Database Technology (EDBT)*, 2016.

[27] S. Lakshminarasimhan, N. Shah, S. Ethier, S. Klasky, R. Latham, R. Ross, and N. F. Samatova, "Compressing the incompressible with isabela: In-situ reduction of spatio-temporal data," in *European Conference on Parallel Processing*. Springer, 2011, pp. 366–379.

[28] E. R. Schendel, Y. Jin, N. Shah, J. Chen, C.-S. Chang, S.-H. Ku, S. Ethier, S. Klasky, R. Latham, R. Ross *et al.*, "Isobar preconditioner for effective and high-throughput lossless data compression," in *IEEE 28th Intl. Conference on Data Engineering*, 2012, pp. 138–149.

[29] J. Jenkins, I. Arkatkar, S. Lakshminarasimhan, D. A. Boyuka II, E. R. Schendel, N. Shah, S. Ethier, C.-S. Chang, J. Chen, H. Kolla *et al.*, "Alacrity: Analytics-driven lossless data compression for rapid in-situ indexing, storing, and querying," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems X*. Springer, 2013, pp. 95–114.

[30] E. Soroush and M. Balazinska, "Time travel in a scientific array database," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 98–109.