TBD-DP: Telco Big Data Visual Analytics with Data Postdiction

Constantinos Costa*, Andreas Charalampous*, Andreas Konstantinidis*‡,
Demetrios Zeinalipour-Yazti* and Mohamed F. Mokbel§

*Department of Computer Science, University of Cyprus, 1678 Nicosia, Cyprus
‡Department of Computer Science & Engineering, Frederick University, 1036 Nicosia, Cyprus
§Qatar Computing Research Institute, HBKU, Qatar and University of Minnesota, Minneapolis, MN 55455, USA
{costa.c, achara28, akonstan, dzeina}@cs.ucy.ac.cy; mmokbel@hbku.edu.qa

***Abstract*—In this demonstration paper, we present the *TBD-DP* operator, which relies on existing Machine Learning (ML) algorithms to abstract Telco Big Data (TBD) into compact models that can be stored and queried when necessary. Our proposed *TBD-DP* operator has the following two conceptual phases: (i) in an offline phase, it utilizes a LSTM-based hierarchical ML algorithm to learn a tree of models (coined *TBD-DP* tree) over time and space; (ii) in an online phase, it uses the *TBD-DP* tree to recover data within a certain accuracy. Our framework also includes visual and declarative interfaces for a variety of telco-specific data exploration tasks. We demonstrate the efficiency of the proposed operator using SPATE, which is a novel TBD visual analytic architecture we have developed. Our demo will enable attendees to interactively explore synthetic antenna signal traces, we will provide, in both visual and SQL mode. In both cases, the performance of the propositions will be quantitatively conveyed to the attendees through dedicated dashboards.**

***Index Terms*—big data, data reduction, visual analytics**

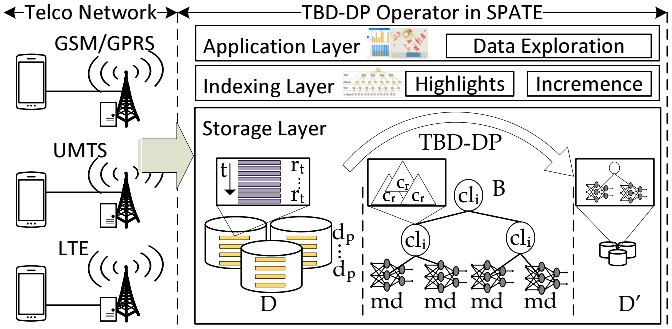

Fig. 1. The TBD-DP operator works on the storage layer of a typical TBD stack and abstracts the incoming data signals (D) into abstract models (md) that are organized in a tree data structure (B).

I. Introduction

Telecommunication companies (telcos) are challenged with the management of petabytes of data per year. For example, a telco in Shenzhen, China serving 10 million users produces 5TB per day [1]. Effectively storing and processing *Telco Big Data (TBD)* workflows can unlock a wide spectrum of challenges, ranging from churn prediction of subscribers, city localization, 5G network optimization/user-experience assessment and road traffic mapping [2]. Even though the acquisition of TBD is instrumental in the success of the above scenarios, Telcos are reaching a point where they are collecting more data than they could possibly exploit. This has the following two implications: (i) it introduces a significant financial burden to store the collected data; and (ii) it imposes a high computational cost for accessing and processing the data.

To this end, *data decaying* [3], [4] (or data rotting) has recently been suggested as a powerful concept to complement traditional data reduction techniques, e.g., sampling, aggregation (OLAP), dimensionality reduction (SVD, DFT), synopsis (sketches) and compression. Data decaying refers to *"the progressive loss of detail in information as data ages with time"*. In data decaying, recent data retains complete resolution, which is practical for operational scenarios that can continue to operate at full data resolution, while older data is either compacted or discarded [2]–[4]. Unfortunately, data decaying currently relies on rather straightforward methodologies, such as rotational decaying (i.e., FIFO) [3], or decaying based on specific queries [2] rather than the complete dataset itself. Our aim is to expand upon these developments to provide more intelligent and generalized decaying operators.

In this demo proposal, we will demonstrate a novel decaying operator for TBD, coined *TBD-DP (Data Postdiction)*, which is technically presented in [5]. *TBD-DP* is implemented over a visual analytic architecture for TBD we have developed, coined SPATE [2]. Unlike data prediction, which aims to make a statement about the future value of some tuple in a TBD store, data postdiction aims to make a statement about the past value of some tuple that does not exist anymore, as it had to be deleted to free up space. *TBD-DP* relies on existing Machine Learning (ML) algorithms to abstract TBD into compact models that can be stored and queried when necessary (see Figure 1). Our proposed *TBD-DP* operator has the following two conceptual phases: (i) in an offline phase, it utilizes a LSTM-based hierarchical ML algorithm to learn a tree of models (coined *TBD-DP* tree) over time and space; (ii) in an online phase, it uses the *TBD-DP* tree to recover data with a certain accuracy. *The demo will enable the audience to see our TBD-DP algorithm in action through an engaging domain-specific demonstration with a visual interface.*

II. Overview of *TBD-DP*

To understand the operational aspects of our proposed *TBD-DP* operator, consider Figure 1, where we show how incoming telco data signals are absorbed by the TBD architecture and stored on high-availability and fast storage (i.e., D). This helps to carry out operational tasks (e.g., alerting services and visual analytics) with full data resolution. Subsequently, in the first phase of *TBD-DP*, we utilize a specialized *Recurrent Neural Network (RNN)* composed of *Long Short Term Memory (LSTM)* units, which has the ability to detect long-term correlations in activity data and the trained model has a small disk space footprint. This enables *TBD-DP* to utilize minimum storage capacity of the decayed data by representing them with LSTM models on the disk media (D') and provide real-time postdictions with high accuracy in a subsequent recovery phase, which will be initiated on-demand (i.e., whenever some high-level operator requests the given data blocks).

Particularly, we express our solution in two internal algorithms, namely, the *Construction* (Data model creation) and the *Recovery* (Data recreation), which capture its core functionality as illustrated in Figure 1.

The *Construction* algorithm can be triggered either by the user, or automatically when the total storage capacity reaches a certain level. In both cases, the data is initially clustered based on the spatial attributes and then ordered based on temporal information. Finally, postdiction models based on the LSTM machine learning approach are generated for each cluster and the real data is decayed by f%. The *Recovery* algorithm utilizes the postdiction models for retrieving the decayed data by adopting a proposed DP-tree based algorithm.

III. Demonstration Scenario

During the demonstration, the attendees will be able to appreciate the efficiency of *TBD-DP* in SPATE, the visualization abstraction and the performance of our propositions.

A. Demo Artifact

We have extended *SPATE* [2], a novel SPARK-based processing architecture with HDFS and an RDBMS for catalog management, in order to integrate our *TBD-DP* operator. The proposed architecture comprises of three layers (see Figure 1), namely Storage Layer, Indexing Layer and Application Layer. The Storage layer uses the *TBD-DP* operator in order to provide the decay methods for the Indexing Layer. The Application Layer uses the index to retrieve the decayed data and consists of a web-based user interface in HTML5/CSS3 along with extensive AngularJS. An illustrative network exploration interface is shown in Figure 2. The *TBD-DP* has been implemented using Tensorflow over HDFS in Python. We have implemented a query sidebar that allows the user to execute a variety of snapshot queries and recurring queries (in the form of a time-machine) for drop calls and downflux/upflux, heatmap statistics and settings. For each query the accuracy of the results will be visualized using fancy charts using the web interface. The hardware stack of our installation resides on our laboratory DMSL datacenter and interaction will be achieved over cable or Wi-Fi using a standard laptop.

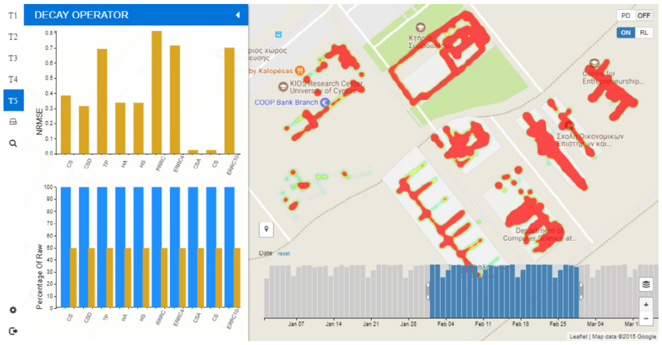

Fig. 2. The *TBD-DP* operator implemented inside the spatio-temporal SPATE architecture. The interface enables users to carry out high resolution visual analytics, without consuming enormous amounts of storage. The savings are quantified numerically with bar charts and visually with heatmaps.

B. Demo Plan

Visual mode: In this mode, the conference attendees will have the opportunity to interactively engage with the *TBD-DP* operator over *SPATE*. We will pre-load a variety of synthetic and web-accessible datasets to the SPATE back-end. The loaded data will capture the structure of real telco data (e.g., open cell tower data, and synthetic CDR and NMS data) and will be very useful to visually show how the *TBD-DP* operator works in real time (i.e., both the indexing of the data but also the querying of it). In order to present the benefits of our propositions to the attendees, we will provide visual cues that will enable the audience to understand the performance benefits (i.e., accuracy, storage, memory and CPU time). The technical comparison of the proposition to state-of-the-art appears in [5].

SQL mode: In this mode, the conference attendees can submit custom SQL queries using auto-complete functionality based on a TBD relational schema we will provide. The *SPATE SQL* interface will allow the attendees to rapidly visualize the result-sets using fancy charts (pie, bar, etc.) and a map-based interface that uses tiles from the OSM service. Our particular aim here will be to describe how the *TBD-DP* residing on the HDFS, will be accessible to all basic block queries, nested queries, joins, aggregates, etc.

References

[1] 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.

[2] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Efficient exploration of telco big data with compression and decaying," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, April 2017, pp. 1332–1343.

[3] M. L. Kersten and L. Sidirourgos, "A database system with amnesia." in *CIDR*, 2017.

[4] 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.

[5] C. Costa, A. Charalampous, A. Konstantinidis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Decaying telco big data with data postdiction," in *19th IEEE International Conference on Mobile Data Management (MDM'18), Aalborg, Denmark, June 25 - June 28 (accepted)*, 2018, p. 10 pages.