SPATE: Compacting and Exploring Telco Big Data

Constantinos Costa, Georgios Chatzimilioudis
[*]Dept. of Computer Science
University of Cyprus
1678 Nicosia, Cyprus
{costa.c, gchatzim, dzeina}@cs.ucy.ac.cy

Demetrios Zeinalipour-Yazti[‡*]
[‡]Max Planck Institute for Informatics
Saarland Informatics Campus
66123 Saarbrücken, Germany
dzeinali@mpi-inf.mpg.de

Mohamed F. Mokbel
Dept. of Computer Sc. & Engr.
University of Minnesota
Minneapolis, MN 55455, USA
mokbel@cs.umn.edu

***Abstract*—In this demonstration paper, we present *SPATE*, an innovative telco big data exploration framework whose objectives are two-fold: (i) minimizing the storage space needed to incrementally retain data *over time*; and (ii) minimizing the response time for spatiotemporal data exploration queries over stored data. Our framework deploys lossless data *compression* to ingest streams of telco big data in the most compact manner retaining full resolution for data exploration tasks. We augment our storage structures with decaying principles that lead to the progressive loss of detail as information gets older. Our framework also includes visual and declarative interfaces for a variety of telco-specific data exploration tasks. We demonstrate SPATE in two modes: (i) Visual Mode, where attendees will be able to interactively explore synthetic telco traces we will provide; and (ii) SQL Mode, where attendees can submit custom SQL queries based on a provided schema.**

Video: https://goo.gl/BNqHFV

I. Introduction

In recent years there has been considerable interest from *telecommunication companies (telcos)* to extract concealed value from their network data. Consider for example a telco in the city of Shenzhen, China, which serves 10 million users. Such a telco is shown to produce 5TB per day [6] (i.e., thousands to millions of records every second). Huang et al. [2] break their 2.26TB per day telco big data 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.

Effectively processing telco big data workflows can unlock a wide spectrum of challenges, ranging from network plan optimization and user experience assessment [2] to city planning and urban engineering [3][5]. Zhang et al. [6] have developed *OceanRT* that improves the visualization capabilities of telco big data by the usage of standard spatio-temporal indexes. Iyer et al. [3] present *CellIQ* to optimize queries, such as “traffic hotspots” and “hand-off sequences with performance problems”, using graph processing. Huang et al. [2] empirically demonstrate that churn prediction performance can be significantly improved with telco big data by integrating both BSS and OSS data. Luo et al. [5] propose a framework to predict user behavior involving more than one million telco users. Prior work is mainly concerned with the analytic exploration of telco big data, while this work optimizes the operational ability of such tasks.

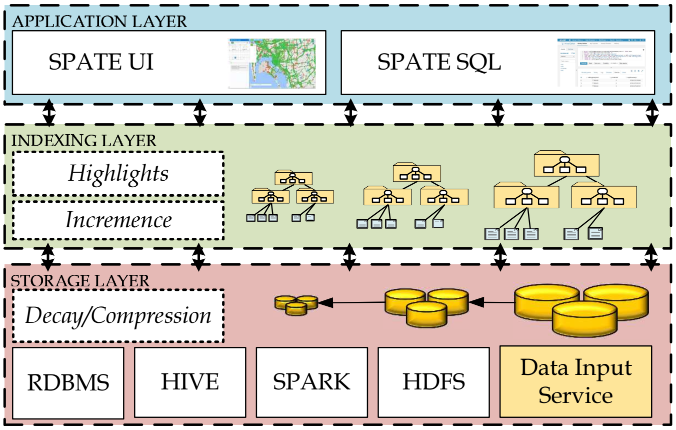


Fig. 1. *SPATE* is an efficient telco big data exploration stack that enables a wide range of smart city applications with a minimal storage cost. It deploys compression, decaying and exploration of the collected data in a unified way.

In this demo we present *SPATE* [1], a framework that uses both lossless data *compression* and lossy data *decaying* 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. It also enables data exploration tasks to retain full resolution over the most important collected data. *Decaying* on the other hand, as suggested in [4], refers to the progressive loss of detail in information as data ages with time until it has disappeared.

SPATE enables data exploration tasks to retain high-level data exploration capabilities for predefined aggregate queries over extremely long time windows, without consuming enormous amounts of storage. It is shown to offer similar performance to the state-of-the-art for telco-specific tasks [1]. Our objective is to minimize the storage costs associated with telco big data exploration tasks, as storage overheads will inevitably lead to the deletion of valuable data, missing in this way the hope to learn valuable insights at the macroscopic scale.

II. Overview of SPATE

We express our solution in three layers (see Figure 1), namely Storage Layer, Indexing Layer and Application Layer.

The *Storage layer* passes newly arrived network snapshots through a lossless compression process storing the results on a replicated big data file system for availability and performance.

This component is responsible for minimizing the required storage space with minimal overhead on the query response time. The intuition is to use compression techniques that yield high compression ratios but at the same time guarantee small decompression times. We particularly use GZIP compression that offers high compression/decompression speeds, with a high compression ratio and maximum compatibility with I/O stream libraries in the big data ecosystem we use. The storage layer is basically only responsible for the leaf pages of the *SPATE* index described in the next layer.

The Indexing Layer uses a multi-resolution spatio-temporal index, which is incremented on the rightmost path with every new data snapshot that arrives (i.e., every 30 minutes). In addition, the component computes interesting event summaries, called "highlights", from data stored in children nodes and stores them at the parent node. For each data exploration query, the internal node that covers the temporal window of the query is accessed, and its highlights are used to answer the query. Finally, this layer is also responsible for the gradual decay of the data. It does so by pruning-off parts of the index tree in using the so called data fungus.

The Application Layer implements the querying module and the *data exploration* interfaces, which receive the data exploration queries in visual or declarative mode and use the index to combine the needed highlights and snapshots to answer the query. *SPATE* is equipped with an easy-to-use map-based web interface layer that hides the complexity of the system through a simple and elegant web interface.

III. Demonstration Scenario

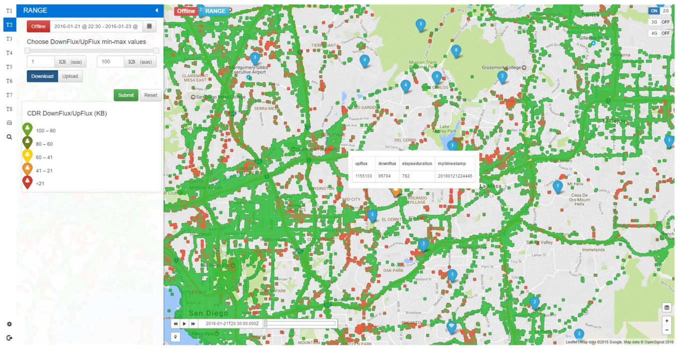
During the demonstration, the attendees will be able to appreciate the key components in SPATE, the visualization abstraction and the performance of our propositions.

A. Demo Artifact

We have implemented a prototype of SPATE using a modern SPARK-based processing architecture with HDFS and an RDBMS for catalog management (see Figure 1). The *SPATE UI (User Interface)* is implemented in HTML5/CSS3 along with extensive AngularJS. An illustrative network exploration interface is shown in Figure 2. We have implemented a query sidebar that allows the user to execute a variety of template queries. The query bar includes snapshot queries and recurring queries (in the form of a time-machine) for drop calls and downflux/upflux, heatmap statistics and settings. Furthermore, quick access buttons are provided so that user is able to choose between the available network modalities (2G, 3G, 4G). The hardware stack of our SPATE installation resides on our laboratory DMSL datacenter and interaction will be achieved over cable or Wi-Fi using a standard laptop, a tablet or smartphones we will bring along at the conference.

B. Demo Plan

Visual mode: In this mode, the conference attendees will have the opportunity to interactively engage with the *SPATE UI*. 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 SPATE compression algorithms work 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., storage, memory and CPU time) and the negligible reduction in query or user-interface response time that we have observed in experiments [1].



Offline 2016-01-21 @ 22:30 - 2016-01-23 @

Choose DownFlux/UpFlux min-max values

1 KB (min) 100 KB (max)

CDR DownFlux/UpFlux (KB)

- 100 – 80
- 80 – 60
- 60 – 41
- 41 – 21
- <21

upflux	downflux	elapseduration	mytimestamp
1155103	95704	782	20160121224445

Fig. 2. SPATE UI: A spatio-temporal telco data exploration user interface we developed on top of Google Maps, which enables combining network models with network measurements encapsulated in the compressed SPATE structures.

SQL mode: In this mode, the conference attendees can submit custom SQL queries using auto-complete functionality based on a telco big data relational schema we will provide. Our hypothesis is that many data engineering researchers and practitioners would feel more comfortable to formulate custom query predicates, as opposed to be limited within the boundaries of well-defined query templates provided by the *SPATE UI*. 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 SPATE structure, residing on the HDFS, will be accessible to all basic block queries, nested queries, joins, aggregates, etc.

Acknowledgments

This work was supported in part by the University of Cyprus, an industrial sponsorship by MTN Cyprus and EU COST Action IC1304. The third author's research is supported by the Alexander von Humboldt-Foundation, Germany. The last author's research is supported by NSF grants IIS-0952977, IIS-1218168, IIS-1525953, CNS-1512877.

References

[1] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, *"Efficient Exploration of Telco Big Data with Compression and Decaying,"* in IEEE ICDE'17.

[2] 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 ACM SIGMOD'15.

[3] A. P. Iyer, L. E. Li, and I. Stoica, *"Celliq: Real-time cellular network analytics at scale,"* in USENIX NSDI'15.

[4] M. L. Kersten, *"Big data space fungus,"* in CIDR'15.

[5] 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. 63:163:30, May 2016.

[6] S. Zhang, Y. Yang, W. Fan, L. Lan, and M. Yuan, *"Oceanrt: Real-time analytics over large temporal data,"* in ACM SIGMOD'14.