Towards Real-Time Road Traffic Analytics using Telco Big Data

Constantinos Costa*, Georgios Chatzimilioudis*, Demetrios Zeinalipour-Yazti‡* and Mohamed F. Mokbel§

* University of Cyprus, Cyprus
‡ Max Planck Institute for Informatics, Germany
§ University of Minnesota, MN, USA

{costa.c, gchatzim, dzeina}@cs.ucy.ac.cy
dzeinali@mpi-inf.mpg.de, mokbel@cs.umn.edu

11th International Workshop on Real-Time Business Intelligence and Analytics, collocated with VLDB 2017, Munich, Germany, August 28, 2017
Motivation

- The expansion of mobile networks and IoT (IP-enabled hardware) have contributed to an explosion of data inside Telecommunication Companies (Telcos)
Motivation

• INRIX reports that congestion costs U.S. drivers nearly **$300 Billion** in 2016, which is an average of **$1400** per driver per year (both direct and indirect costs)

<table>
<thead>
<tr>
<th>Rank</th>
<th>City / Large Urban Area</th>
<th>2016 Peak Hours Spent</th>
<th>Total Cost Per Driver in 2016</th>
<th>Total Cost to the City in 2016 (based on city population size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Los Angeles, CA</td>
<td>104</td>
<td>$ 2,408</td>
<td>$9.7bn</td>
</tr>
<tr>
<td>2</td>
<td>New York, NY</td>
<td>89</td>
<td>$ 2,533</td>
<td>$16.9bn</td>
</tr>
<tr>
<td>3</td>
<td>San Francisco, CA</td>
<td>83</td>
<td>$ 1,996</td>
<td>$2.5bn</td>
</tr>
<tr>
<td>4</td>
<td>Atlanta, GA</td>
<td>71</td>
<td>$ 1,861</td>
<td>$3.1bn</td>
</tr>
<tr>
<td>5</td>
<td>Miami, FL</td>
<td>65</td>
<td>$ 1,762</td>
<td>$3.6bn</td>
</tr>
</tbody>
</table>

Telco Big Data (TBD)

- **Telco Data**: Traditional source for OLAP Data Warehouses and Analytics.
  - e.g., Accounting, Billing, Session data.
  - **Problem**: Inadequate data resolution to address biggest challenges:
    - e.g., churn prediction, 5G network optimization, user-experience assessment, traffic mapping.

- **Telco Big Data (TBD)**: Velocity data generated at the cell towers.
  - e.g., signal strength, call drops, bandwidth measurements.
  - **Size**: 5TBs/day for 10M clients (i.e., 2PB/year).
Challenge

• Road traffic analytic and prediction systems are not sufficient.
  – traffic modeling and prediction is limited to navigation enterprises utilizing **crowdsourcing**
  – the location **privacy** boundaries of users could be compromise by third-party mobile applications
  – the **cost** of continuously monitoring the traffic is very high

• **Our Approach:** Introduce a **Traffic-TBD** (Traffic Telco Big Data) architecture that goes beyond crowdsourcing retaining the user privacy and minimizing the cost using existing resources.
Presentation Outline

- Introduction
- **Background**
- Traffic-TBD Architecture
  - Data Layer
  - Processing Layer
  - Application Layer
- Experiments
- Conclusions
Background: Traffic analytics

• First Era (1E) traffic mapping
  – dedicated hardware
  • magnetic loop detectors or traffic cameras mounted on traffic lamps or road-sides

• Second Era (2E) traffic mapping
  – handheld-based and probe vehicles
  • Taxis or Buses were equipped with a smartphone and a dedicated app for location updates
  • MIT Cartel (http://cartel.csail.mit.edu/doku.php)
Background: Traffic analytics

• Second+ Era (2.5E) traffic mapping
  – opportunistic crowdsourcing

• Third Era (3E) traffic mapping
  – Traffic-TBD: TBD data to generate the traffic mapping both accurately, without additional hardware and additional costs.
  – E.g., TomTom MyDrive
• Radio & Core Network
  – Generate Streams of IP packets from the GSM/GPRS (2G), UMTS (3G), LTE (4G) antennas.
• **Telco Big Data (TBD)**
  - Business Supporting Systems (BSS) data
    - Tens of **GB/day** for a xM+ city
  - Operating Supporting Systems (OSS) data
    - Tens of **TB/day** for a xM+ city
Presentation Outline

• Introduction
• Background
• **Traffic-TBD Architecture**
  – Data Layer
  – Processing Layer
  – Application Layer
• Experiments
• Conclusions
Traffic-TBD Architecture

- The Traffic-TBD architecture is based on SPATE* and has 3 layers:
  - Data layer
    - aggregate and store data
    - provide the access methods
  - Processing layer
    - minimizing the query response time for data analytic queries
  - Application layer
    - enhance user experience

Data Layer: Components

• Storage component:
  – minimizes storage

• Indexing component:
  – maintains and uses a multi-resolution spatiotemporal index
Storage Component: Compression

- **Compression** refers to the encoding of data using fewer bits than the original representation.

- **Benefits:**
  - Shifts the resource bottlenecks from **storage**- and **network**-I/O to **CPU** (cycles increasing faster).
  - Save enormous **amounts of storage** and I/O on subsequent analytic tasks!

- **Desiderata:** Given a setting where TBD arrives **periodically** in batches, we want to:
  - minimize the **space** needed to store/archive data
  - minimize response time for spatiotemporal data exploration queries
Storage Component: Compression

- We analyzed an anonymized real TBD dataset to understand what compression ratios can be achieved?
  - **Dataset:** 1 week, 1.7M CDR, 21M NMS, 300K users
  - **Tool:** Shannon’s entropy (typical measure for the maximum compressibility of a dataset)
  - **Observation:** TBD contains significant redundancy (Entropy close to 0)!

![CDR: Entropy of Attributes](image1)

![NMS: Entropy of Attributes](image2)
## Storage Component: Compression

<table>
<thead>
<tr>
<th>Libraries\ Objectives</th>
<th>GZIP</th>
<th>7z</th>
<th>SNAPPY</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression Ratio ($r_c$)</td>
<td>9.06</td>
<td>11.75</td>
<td>4.94</td>
<td>9.72</td>
</tr>
<tr>
<td>Compression Time ($T_{c1}$) in sec</td>
<td>21.37</td>
<td>20.99</td>
<td>21.39</td>
<td>21.07</td>
</tr>
<tr>
<td>Decompr. Time ($T_{c2}$) in sec</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
<td>0.11</td>
</tr>
</tbody>
</table>

* GZIP also readily available by most Stream I/O libraries & application layer (e.g., HIVE, HUE)
Indexing Component: Modules

- Our spatio-temporal index has 4 levels of temporal resolutions
  - epoch (30 minutes), day, month, year.
- **A) Incremence Module**
  - Add compressed snapshots on the right-most path
B) Highlights Module
- materialized views to long-standing queries of users (e.g., the drop-call counters, bandwidth statistics, etc.)
- essential for multi-granular visual analytics!
Processing Layer: Components

• Map Matching:
  – generate traffic points
    • match them to the road network

• Privacy:
  – Location data has high privacy requirement
    • it is easy to infer user activity from trajectories

• Visualization:
  – efficiently pack and ship data to the UI

• Mining:
  – interesting incidents through the TBD
Application Layer: Queries

• **A1. Hot Spot Analysis:**
  – Detect and visualize the most congested areas around a location.

• **A2. Navigation Monitoring:**
  – Calculate and visualize the best route based on the travel time and traffic queue length.

• **A3. Travel time:**
  – Calculate aggregated time loss due to congestion.

• **A4. Travel speed:**
  – Determine average speed based on the traffic flow.
Application Layer: Queries

• **A5. Accessibility analysis:**
  – Compute the time that is needed to reach a specific location at a specific time slot considering the road traffic.

• **A6. Incident detection:**
  – Detects any incidents due to traffic.

• **A7. Before-after analysis:**
  – Measure the traffic flow after a change in the road network in order to be evaluated from the traffic administration.

• **A8. Travel behaviors:**
  – Identify travel behaviors throughout a region.
Application Layer: Components

- **Interactive Traffic Analytics:**
  - receives a data analytic query and uses the index to combine the needed highlights to answer the query

- **Alert Service:**
  - is a background service that can provide real-time and predicted traffic alerts combining public APIs with the TBD

- **API:**
  - makes the generated traffic insights available for public use.
Presentation Outline

• Introduction
• Background
• Traffic-TBD Architecture
  – Data Layer
  – Processing Layer
  – Application Layer
• Experiments
• Conclusions
Experimental Methodology

• To evaluate the *DATA layer*, we have implemented a *trace-driven* testbed:

• Compared Frameworks:
  – **RAW**: stores the snapshot on disk without compression or indexing
  – **SHAHED**: stores the snapshot using a quad-tree index (part of the offline analytic SpatialHadoop 2.4 framework @ ICDE’15).
  – **SPATE**: the underlying framework used in this work.

• Metrics:
  – **Ingestion Time (sec)**: time cost to store a 30-min TBD snapshot
  – **Space (MB)**: space cost to store a 30-min TBD snapshot
  – **Response Time (sec)**: time to answer a query.
Experimental Testbed

- **TBD Operating System Stack**
  - **Datacenter**: VMWare ESXi 5.0.0 Hosts
  - **VMs**: 4 Ubuntu 14.04 server images, each featuring: 8GB of RAM with 2 virtual CPUs (2.40GHz)
  - **Storage Element**: Slow 7.2K RPM RAID-5 SAS, 6 Gbps disks. Each disk formatted in VMFS 5.54 (1MB block)

- **TBD Framework Stack**
  - Hadoop Distributed File System (HDFS) v2.5.2
  - Apache Hive 2.0 (online querying)
  - Apache Spark 1.6.0 (offline data processing)
    - Individual Scala programs submitted directly to the Spark computation master
Ingestion Time and Space

- **Ingestion time (left)** and **Space (right)** of methods over different time windows within a day.

- **Observation:** SPATE 1.25x slower in ingestion but needs 1 order of magnitude **less storage space**!
  - Ingestions are separated by 30 minute windows, so this is acceptable.
Data Exploration Tasks

**Observation:**
- **CPU intensive tasks:** SPATE = RAW = SHAHED even though we use 1 order of magnitude **less** disk space!
Conclusions

• Conclusions:
  – Traffic-TBD architecture aims to provide:
    • micro-level traffic modeling and prediction
    • location privacy boundaries of users inside their mobile network
    • availability with minimal costs and using existing infrastructure

• Future Work:
  – map-matching of trajectories
  – privacy-preserving trajectory processing
Efficient Exploration of Telco Big Data with Compression and Decaying

Thanks! Questions?

Constantinos Costa*, Georgios Chatzimilioudis*, Demetrios Zeinalipour-Yazti‡* and Mohamed F. Mokbel§

* University of Cyprus, Cyprus
‡ Max Planck Institute for Informatics, Germany
§ University of Minnesota, MN, USA
{costa.c, gchatzim, dzeina}@cs.ucy.ac.cy
dzeinali@mpi-inf.mpg.de, mokbel@cs.umn.edu

Eleventh International Workshop on Real-Time Business Intelligence and Analytics
Munich, Germany, August 28, 2017