Deterministic Real-Time Analytics of Geospatial Data Streams through ScaleGate Objects

Vincenzo Gulisano, Yiannis Nikolakopoulos, Ivan Walulya, Marina Papatriantafilou and Philippas Tsigas

NOMINATED FOR DEBS 2015 GRAND CHALLENGE AWARD

DEBS 2015 GRAND CHALLENGE

Analize taxi trip reports from NYC and compute:

- Top-10 most frequent routes in the last 30 minutes.
- Top-10 most profitable areas based on the median fare and tip (during the last 15 minutes) and number of empty taxis (during the last 30 minutes)

- Enables concurrent and in-order deterministic processing of ready tuples in data streaming
- Lock-free linearizable implementation enables determinism and full fine-grain concurrency

OUR APPROACH AND NOVELTY

Scale up, then scale out!

- New pioneering data structures with appropriate API and concurrent implementations, enabling
- Enhanced Parallel and Distributed Stream Processing Engine’s analysis

AT THE CORE:

Key data structures maintained by the Processing Units

Query 1: Top-K most frequent routes
- Order events’ occurrences using routes as unique key
- Provide Top-K counts in O(1) time.
- All operations with O(1) time complexity on average
- Worst case: linear in hashtable size

Query 2: Top-K most profitable areas
- Calculate median over a sliding window
- O(logN) w.h.p. on new tuple
- O(1) on average on expired tuples
- Maintain PriorityQueue for profitable areas

IMPLEMENTATION

PERFORMANCE

Applicability

Virtual machine with 4 cores, running on an Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz
(cache size: 6144 KB)
Throughput: 110,000 tuples/second
Latency: 46 milliseconds