Ingest-time transformations in Time Series Database

Henggang Cui

Kimberly Keeton, Indrajit Roy, Krishnamurthy Viswanathan, Haris Volos (HP Labs)
Gregory R. Ganger

PARALLEL DATA LABORATORY
Carnegie Mellon University
Time Series Data Analytics

Time Series Database

Analytical queries

Metric M1

Metric M2

correlation

Value

Time

http://www.pdl.cmu.edu/

Henggang Cui © October 14
Goals and an example use case

Power usage monitors

Time Series Database

(1) Interactive queries: need sub-second latency

Find most correlated range in 
{monitor0, yesterday} 
with {monitor1, 3am to 5am},

(2) Queries on both recent data and historical data
Approaches

- Streaming queries (e.g., Storm)
  - Only for pre-determined queries
  - No interactive queries, no revisiting history
- Systems for queries on recent data (e.g., Scuba)
  - Not for historical data
- Sampling approaches (e.g., BlinkDB)
  - Can miss too much information (e.g., spikes)

- Our approach: ingest-time transformations
  - Reduce work to be done at query-time
Ingest-time transformations

Ingest-time processing

- Raw Data
  - Ingest
  - Transform
  - Transformation-1
    - Transformed data
  - Transformation-2

Query-time processing

- Example transforms
  - Fourier, wavelet, ARMA coefficients
Example: wavelet transformation

- Compact representation of signal
  - Similar to Fourier transform
  - But better at capturing spikes
- Can be approximated with few coefficients
  - Much smaller than raw data
  - With just a small amount of error
Example: wavelet transformation

Ingested raw data

Wavelet transform,
4096 points per window,
with 10% error

The large wavelet coefficients
that satisfy the error bound
5% the size of raw data
-> more efficient queries
Correlation using wavelet coeffs

- Query latency vs. error bound

![Graph showing query latency vs. error bound for raw data and wavelet with 10% and 20% error.](http://www.pdl.cmu.edu/)
Using pre-computed transforms

+ Reduced latency
  - Less data to be retrieved and processed

- Can only be used when
  - Query can tolerate the bounded error
  - Time ranges are multiples of the window size

  - Idea: keep transforms of multiple data resolutions
    - Can increase likelihood that query can use one
Multiple data resolutions

- Transforms can be applied
  - (1) With different error bounds
  - (2) In different window sizes

More compact transformed data.
Fewer queries can use

Window size

Error bound
A complete use case example

Transforms defined:
- Wavelet(window=1min, error=10%)
- Wavelet(window=1hour, error=10%)
- Wavelet(window=1hour, error=20%)

Find most correlated range in {monitor0, yesterday} with {monitor1, 3am to 5am}, with data error bound 10%

Correlation calculated using transform Wavelet(window=1hour, error=10%)

Power usage monitors
Continuing work

• Prototype implemented in LazyBase
  • Supported transforms:
    – Fourier, Wavelet, ARMA coefficients
• Exploring additional transformations
  • And query optimization challenges

• Looking for more …
  • Use cases
  • Datasets
References


