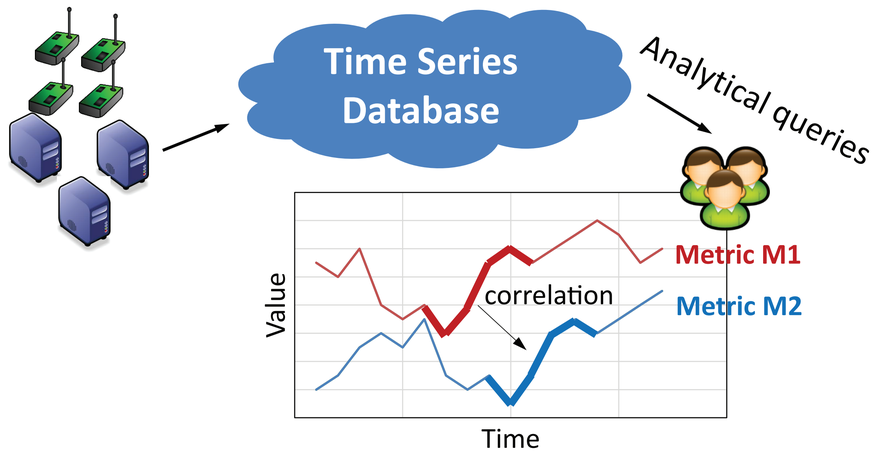
Using Data Transformations for Low-latency Time Series Analysis

Henggang Cui, Kimberly Keeton (HP Labs), Indrajit Roy (HP Labs), Krishnamurthy Viswanathan (HP Labs), Greg Ganger

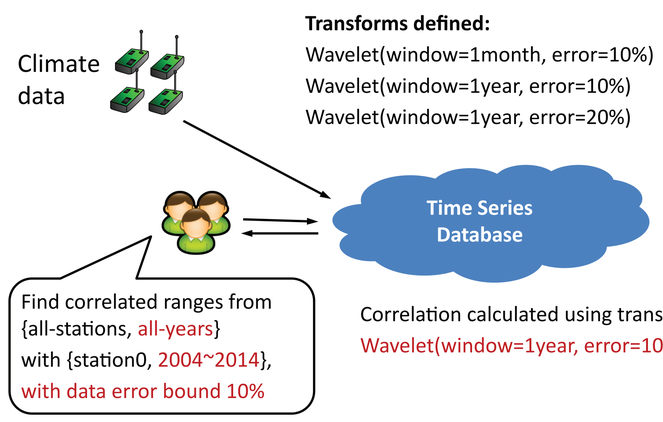
Time Series Data Analytics

- Time series data
  - › Sensors, cluster performance counters, etc.
- Analytical queries
  - › E.g., find data correlated to another range of data of data
- Our goals
  - › Interactive queries: need sub-second latency
  - › Queries on both recent data and historical data

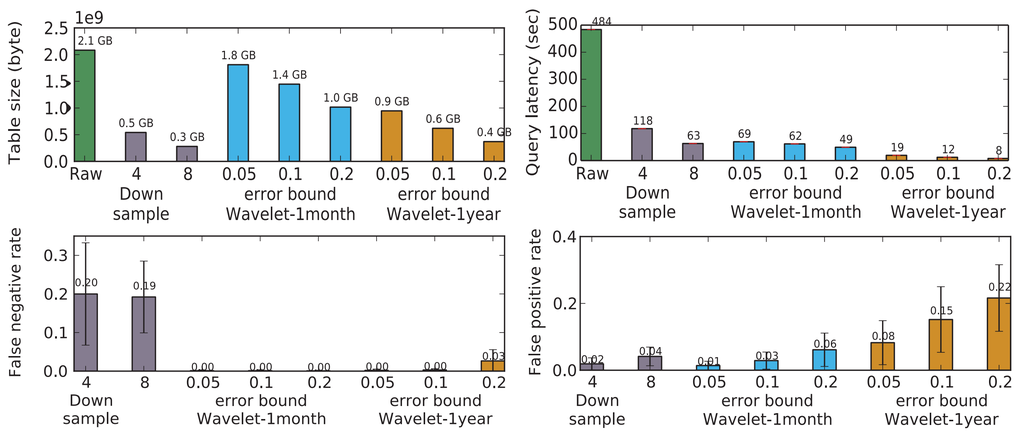


Example Use Case: Correlation Search

- Find data ranges with correlation larger than a threshold
- Can be approximated with frequency domain transformations (e.g., wavelet)
  - › Bounded error and much smaller than raw data



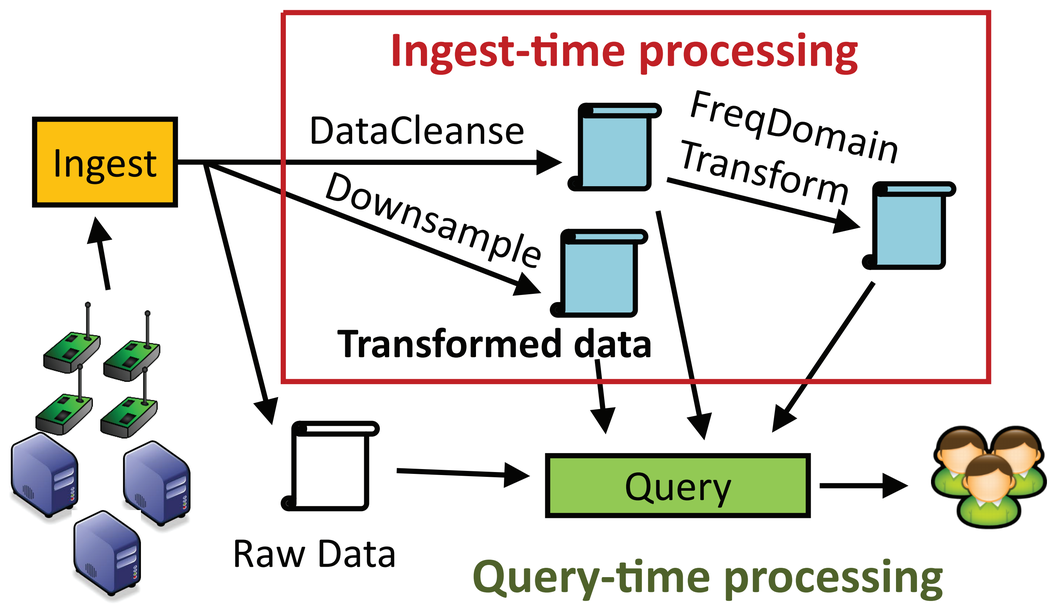
- Experiment: find correlated timeseries windows
  - › Dataset: climatic data with 350 million data points



- Some takeaway observations
  - › Only 1.7% of the baseline latency with false negative/positive rates 3%/22%
  - › 4% ingestion overhead when doing six wavelet and two downsample transformations

Approach: Data Transformations

- User-defined transformations on ingested data
  - › Transformed data and raw data both kept
- Each query uses most efficient option



- Ingest-time processing
  - › Transformations based on user-defined windows
    - E.g., every hour of data collected from one sensor
  - › Chained transformations
    - E.g., data cleansing before others
  - › Keeping multiple versions of transformed data
    - E.g., with different window granularities and error bounds
- Query-time processing
  - › Automatic transformed data selection
    - Based on user-defined utility functions
  - › Translate queries to use transformed data

Other Use Cases

- Anomaly detection
  - › Calculate and store ARMA residuals at ingest-time
  - › 2.8% of the baseline query latency
- Event occurrence monitoring
  - › Use count-min sketch to summarize count info compactly
  - › 40,000x lower latency with 12% error
- Answering arbitrary queries
  - › Reconstruct original data from wavelet transformed data
  - › Provide data in 1/50th the time with 13% error

Carnegie Mellon University

Parallel Data Laboratory

PDLSVD15