RecovDB: accurate and efficient missing blocks recovery for large time series

Ines Arous¹, Mourad Khayati¹, Philippe Cudré-Mauroux¹
Ying Zhang², Martin Kersten², Svetlin Stalinlov²

¹{firstname.lastname}@unifr.ch
²{firstname.lastname}@monetdbsolutions.com

GOAL AND CONTRIBUTIONS

Motivation: Real-world time series (sensor) data often contain missing values. Missing values are harmful to upper-level time series analytics.

Goal: An efficient and accurate DB system to recover missing values in time series.

Contribution: A system called RecovDB with:
- Parameter-free recovery
- Correlation-aware recovery
- Full-fledged DBMS (MonetDB) support

RECOVERY CHALLENGES

- Long and (linearly) correlated time series.
- Large missing blocks (up to 80% of consecutive observations are missing per time series).
- Multiple incomplete time series.
- Integrate the recovery in the MonetDB system.

RECOVERY GUI

Menu

1. Data normalization options.
2. Select one or multiple TS from which to drop blocks of missing values.
3. Select a recovery termination threshold.
4. Select percentage of additional missing values.
5. Use PHP-based recovery instead of UDF.
6. Recover the missing blocks with the selected setup.

EMPIRICAL EVALUATION

Efficiency

- RecovDB is up to 10000x and 10x faster than BayesDB⁴ and ImputeDB⁵ respectively.

To achieve this performance, RecovDB exploits the analytical power of MonetDB to handle the data management and pre-/post-processing.

Accuracy

- RecovDB outperforms the STOA using RMSE, MSE and MAE metrics.
- RecovDB is up to 66% and 87% more accurate than BayesDB and ImputeDB, respectively.
- The accuracy of RecovDB is steady with increasing % of missing values.

DEMO SCENARIOS

Scenario 1 Recover multiple incomplete time series at one.
Scenario 2 Increase the size of the missing block and of the data.
Scenario 3 Compare RecovDB against STOA recovery DB systems (i.e., BayesDB and ImputeDB).

CONCLUSIONS

- We present RecovDB which recovers large missing blocks in multiple time series.
- RecovDB leverages the correlation across time series during the recovery.
- RecovDB outperforms STOA in both efficiency and accuracy when increasing i) the length/number of time series and ii) the size of the missing blocks.

‡Saad, F. and Mansinghka, V.K. A probabilistic programming approach to probabilistic data analysis. NIPS, 2016
§Cambones, J., Feser, J.K., Smith, M.J. and Madden, S. Query optimization for dynamic imputation VLDB, 2017