Contribution ID: **150** Type: **Paper presentation**

Information Retention in Trimmed Datasets

Thursday 18 September 2025 12:10 (15 minutes)

The structure and usage scenarios of a software package for trimming datasets while having minimum information loss are described. Several information models applied to a large dataset generated by an enterprise information system are analyzed. Different strategies and procedures are compared to obtain the best compromise between computing time and information retention. A set of data profiling tests is presented with the purpose of detecting anomalies such as data flooding. The results show that a block trimming strategy allows the preservation of most of the information while speeding up the computation by one or more orders of magnitude. The software automatically detects the optimum trimming level associated with the model, allowing autonomous real-time control of large datasets.

Author: MANOLACHE, Florin Bogdan (Carnegie Mellon University)

Co-authors: RUSU, Octavian (Alexandru Ioan Cuza University, Iasi, Romania); SU, Xinran (Carnegie Mellon University); WANG, Zhilan (Carnegie Mellon University)

Presenter: MANOLACHE, Florin Bogdan (Carnegie Mellon University)

Session Classification: High Performance Computing in Science

Track Classification: Grid, Cloud & High Performance Computing in Science