CLUE — Clustering-Based Load Understanding and Exploration: Summarizing High-Dimensional Electricity Grid Data for Scenario Analysis

Linus Magnusson
gusmagliag@student.gu.se
Gothenburg University
Göteborg, Sweden

Rasmus Thorsson
gushthkhra@student.gu.se
Gothenburg University
Göteborg, Sweden

Quang Vinh Ngo
vinhq@chalmers.se
Chalmers University of Technology
and Gothenburg University
Göteborg, Sweden

Marina Papatriantafilou
ptrianta@chalmers.se
Chalmers University of Technology
and Gothenburg University
Göteborg, Sweden

Joris van Rooij
joris.vanrooij@goteborgenergi.se
Göteborg Energi AB
Göteborg, Sweden

Mihail Chigrichenko
mihail.chigrichenko@goteborgenergi.se
Göteborg Energi Nät AB
Göteborg, Sweden

Abstract

Modern electricity grids generate large volumes of high-dimensional time series data through Advanced Metering Infrastructure (AMI). While this data contains valuable operational insights, its scale and complexity pose significant analytical challenges, including computational constraints, domain knowledge gaps, and the need for targeted exploration. We present an integrated toolchain for data summarization and clustering-based analysis that bridges this gap, giving grid operators practical capabilities to extract actionable insights from complex measurements without requiring advanced algorithmic expertise.

Our toolchain integrates streaming data processing, efficient exploration techniques, configurable and extensible feature engineering, and pattern identification components. This infrastructure enables computationally efficient high-dimensional data processing while maintaining the analytical depth necessary for operational decision-making.

In this article we describe a work in progress and showcase electricity consumption behavior analysis as one example application. The underlying data processing infrastructure supports various analytical tasks across multiple domains.

Keywords

Toolchain, Summarization, Data Analysis, Clustering, DBSCAN, K-means, Timeseries, Electricity, Feature Extraction

ACM Reference Format:

Linus Magnusson, Rasmus Thorsson, Quang Vinh Ngo, Marina Papatriantafilou, Joris van Rooij, and Mihail Chigrichenko. 2018. CLUE — Clustering-Based Load Understanding and Exploration: Summarizing High-Dimensional Electricity Grid Data for Scenario Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In modern electricity grids, the addition of Advanced Metering Infrastructure (AMI) [10] has increased the data collected. This data contains valuable operational insights but presents significant analytical challenges due to its volume and dimensionality. Each consumer's data forms a time series that can be divided into high-dimensional vectors representing temporal usage patterns over a time period as a windowed time series vector. Furthermore, consumption behaviors change over time due to seasonal factors, policy changes, and economic conditions, necessitating sophisticated tracking and analysis methods.

Traditional analytical approaches struggle with these challenges due to computational limitations, data transformation complexities, and the difficulty of extracting actionable insights from high-dimensional raw measurements. When analyzing electricity consumption patterns, looking at raw time series data alone is inefficient both computationally and analytically—it requires substantial processing resources while often failing to reveal specific behavioral patterns relevant to grid management. Moreover, practical interpretation typically requires domain expertise from grid operators, and manual or non-systematic review of such complex data is prohibitively time-consuming, making it impractical to derive actionable insights without appropriate tools regularly.

Our research presents an integrated, configurable, and extensible toolchain that combines streaming data processing, efficient clustering techniques, and feature engineering to address these challenges. The tool chain enables exploration and summarization of high-dimensional consumption data that while maintaining the ability to answer specific grid management questions. For example, operators can identify customers who contribute significantly to peak demand or exhibit rapid consumption increases during particular periods, patterns that are computationally expensive to detect in raw data.

2 Background

Time series clustering presents unique challenges due to high dimensionality, temporal ordering, and complex similarity measures [4]. In real-world electricity grid applications, there is no "one-size-fits-all" solution to data analysis problems. The optimal clustering parameters, algorithms, distance metrics, and even the definition of a meaningful pattern vary significantly depending on the question being addressed. For example, identifying peak demand contributors requires different analytical approaches than detecting seasonal consumption shifts or rapid usage changes.

Density-based clustering algorithms, particularly DBSCAN [6], have shown promise for time series data due to their ability to identify clusters of arbitrary shapes and handle noise. However, standard DBSCAN implementations have severe limitations for large, high-dimensional datasets.

Approximation techniques like Locality-Sensitive Hashing (LSH) address computational challenges in high-dimensional clustering. IP.LSH.DBSCAN [8] integrates LSH directly into the clustering process, creating an efficient parallel algorithm that scales effectively with dataset size and dimensionality. Keramatian et al. [8] demonstrated that IP.LSH.DBSCAN achieves significant speedups compared to traditional DBSCAN while maintaining high-quality clustering results, making it particularly suitable for an analytical toolchain's exploration and parameter optimization components.

The data from the AMI smart meters arrives in a streaming fashion. While research on handling streaming electricity data is active [12] and direct clustering of streaming data has been proposed in the past (cf e.g. [1, 7]), our toolchain supports multiple approaches to temporal data processing. For electricity consumption analysis, temporal contexts (such as daily or weekly patterns) are often valuable, but the toolchain doesn't enforce any particular temporal partitioning. This flexibility enables analysts to select the most appropriate temporal representation for their analysis while preserving the relationships critical for understanding electricity consumption patterns.

Feature engineering approaches extract relevant characteristics from time series data to create more manageable and interpretable representations [2]. Common features include statistical measures (mean, variance), temporal patterns (peaks, load factor), and domain-specific indicators.

3 Challenges

Clustering high-dimensional data poses challenges as dimensionality is a factor in the computational cost of clustering. Furthermore, high-dimensional data tends to behave differently from low-dimensional data, often referred to as the curse of dimensionality [13]. An example of this can be seen for the Euclidean between high-dimensional points. The Euclidean distance between pairs of high-dimensional points tends to converge to the same value, leading to difficulties when clustering high-dimensional data using euclidean distance as the distance metric.

As clustering high-dimensional data is an expensive and often a time-consuming process, there is sometimes a need for trade-offs between accuracy and speed. Ideal clustering parameters must be established to effectively determine outliers in the raw data. This can be done via parameter exploration, which requires the clustering algorithm to be run repeatedly on the same data with varying parameters to find the best fit. This necessitates the algorithm be fast even when applied to a large amount of high-dimensional data.

Another challenge is the expertise gap between domain experts of different domains in this cross-disciplinary field. Electricity grid operators possess crucial knowledge about significant patterns, but they are most often not experienced with algorithms and algorithmic implementations of relevant methods. On the other hand, data engineers can create sophisticated pipelines but typically lack the electricity/energy domain expertise to identify relevant patterns. This gap creates challenges in parameter selection, result interpretation, and iterative refinement, emphasizing the need for a toolchain accessible to domain experts.

User behaviors are difficult to pinpoint from the raw time series data, as the raw data contains a large amount of behavioral information. To convert the raw data into behavioral information that a human can interpret, there is a need to transform the data. This can be done by extracting features from the raw data to focus on a specific behavior question. Selecting features to answer a behavior question is done manually and requires an understanding of the domain-specific interactions between features in order to get relevant information from the results.

4 Toolchain

Our work proposes a flexible, integrated toolchain that connects specialized components, enabling efficient exploration and targeted analysis of electricity grid data, as illustrated in Figure 1. The toolchain integrates data summarization principles with streaming time series clustering through four interconnected stages:

Pre-processing: Conversion of the AMI time series data into a windowed time series vector for each logical input stream. This is done primarily via a pipeline of Apache Kafka [3] and Apache Flink [5]. Apache Kafka is a distributed message broker that efficiently handles high-volume streaming data with fault tolerance. Apache Flink provides a scalable stream processing framework with precise time- and state management capabilities. These technologies enable the reliable ingestion and transformation of continuous meter data streams. This stage constructs the raw time series vectors in preparation for feature extraction.

Feature Extraction and Filtering: Extracting targeted features from the raw time series data to capture the specialized behaviors of the consumers. The main behavior questions include, e.g., peak contribution during critical hours and consumption variability, expressed through various aggregates. This type of data aggregation allows us to deconstruct the overarching behavior in the raw time series into more interpretable sub-behaviors. Furthermore, the raw data is also split using clustering with broad parameters to filter out outliers.

Clustering: Implementation of clustering techniques suited for clustering of features extracted from the raw time series data. These techniques depend on the features in question and can vary depending on feature selection. This clustering aims to identify large enough groups of customers to be relevant for potential behavior analysis.

Behavior Analysis: Tracking the evolution of these question-specific consumer behaviors over time allows us to observe changes when a relevant event occurs or to identify seasonal patterns and trends.

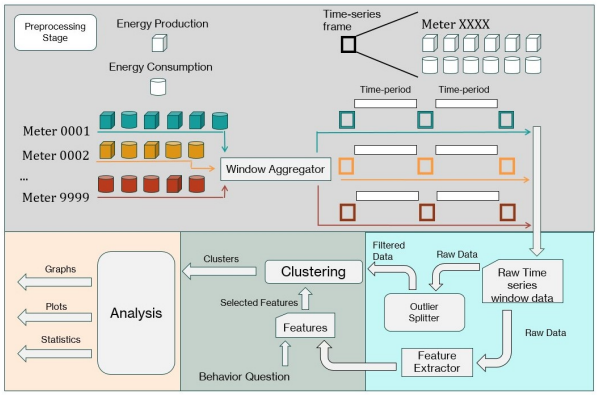

Figure 1: Visualization of the four stages of our proposed system: The pre-processing stage, the feature extraction stage, the clustering stage, and the behavior analysis stage.

As seen in Figure 1, the pre-processing stage aggregates the incoming electricity consumption and production data into time series frames. This is done per meter using a pre-determined window size set to 24 hours for parts of this work. The window frames for the meters for a specific time window are sent in tandem to the outlier splitter to remove outliers from the raw time series vectors, and to the feature extrator for feature extraction. The outliers are split from the main cluster using density-based clustering and removed from the data as they would negatively impact the feature clustering ability further down the pipeline. The remainder of the data is then sent to the clustering stage.

Features are then extracted from the raw data. A subset of these features is selected based on the behavior question for clustering. The clustering stage clusters the selected features using a clustering algorithm. The clustering results are forwarded to the analysis stage, where they are compared to clustering results from previous time periods and eventually converted to graphs, plots, and statistics for expert analysis.

5 Preliminary Findings

This section describes our initial evaluation of the CLUE toolchain using real-world electricity consumption data from Göteborg Energi. We analyze both the computational performance of our approach and the quality of insights generated.

Metrics of interest

The evaluation metrics are latency and accuracy, defined respectively as the time it takes for the toolchain to generate results from the end of the pre-processing stage and the accuracy of the results with respect to some ground truth.

The clustering latency and computational efficiency in the outlier splitting and clustering stages is expected to impact toolchain latency significantly. To have a responsive system with parameter exploration functionality, these two stages must be as efficient as possible. As high-dimensional data tends to be computationally costly to cluster, comparisons concern the usage of regular DBSCAN [6] and IP.LSH.DBSCAN [8].

Some ground truth must be established for the clustering stage to measure the system's accuracy. This can be done by employing expert knowledge about the dataset. By clustering using a dataset of small-to-medium-sized businesses, we can establish the distribution in the clusters of these businesses by SNI categories [11] known to Göteborg Energi. SNI is a national classification system of businesses by categories with varying levels of sub-categories. With these categories, we can evaluate whether the result of the clustering stage is representative of the expected behavior of these businesses. The hopeful result is that certain business categories are found almost exclusively within specific clusters.

Our toolchain's data summarization process addresses the challenges of high-dimensional, high-volume, and high-velocity data by reducing raw measurements to meaningful features. This reduction involves trade-offs between computational efficiency (time and memory usage) and analytical precision.

We applied our toolchain to electricity consumption data from Göteborg Energi, analyzing patterns from approximately 7500 meters of small—to medium-sized businesses. Our initial findings demonstrate several advantages of our approach.

Computational Efficiency

The rapid exploration phase using IP.LSH.DBSCAN processed high-dimensional consumption patterns dramatically faster than standard DBSCAN implementations, making interactive parameter exploration feasible. We evaluated performance on two datasets: Dataset 1 consists of hourly measurements from 16th January 2024, covering approximately 7500 small-to-medium businesses (24-dimensional vectors). Dataset 2 contains data from 5000 meters from the 7th of March 2025, with readings at 15-minute intervals (96-dimensional vectors). Both datasets cover a whole 24-hour period. Table 1 shows the performance comparison of DBSCAN to IP.LSH.DBSCAN across these two datasets.

Dataset	Algorithm	Threads	Avg. Time (s)	Speedup
1	DBSCAN	1	44.055	1.0x
1	IP.LSH.DBSCAN	1	0.504	87.2x
1	IP.LSH.DBSCAN	4	0.508	86.6x
2	DBSCAN	1	124.91	1.0x
2	IP.LSH.DBSCAN	1	3.379	36.9x
2	IP.LSH.DBSCAN	4	2.579	48.1x

Table 1: Computational performance comparison across different datasets

These results demonstrate the substantial efficiency gains of IP.LSH.DBSCAN across different datasets. The implementation of IP.LSH.DBSCAN was tested both sequentially and with utilizing parallel computation with four threads, providing speedups ranging from 49x to 87x compared to traditional DBSCAN. In the case of Dataset 1, the parallelism of IP.LSH.DBSCAN provided no improved results when compared to the sequential version. For Dataset 2 the

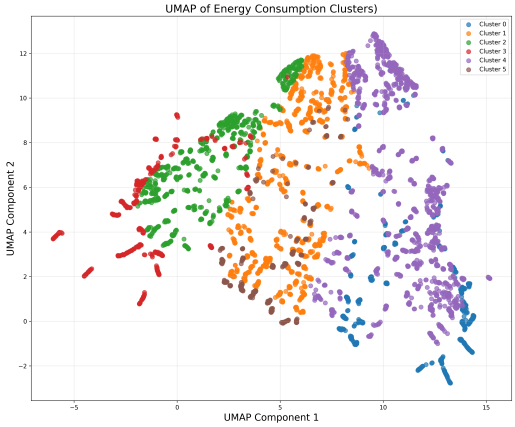

Figure 2: UMAP visualization of clusters derived from features: Peak-to-average ratio, Top peak hour, Number of Significant Peaks, Peak Width.

parallel version shows a clear improvement over the sequential version, suggesting that the dimensionality is more relevant than the size of the dataset in this case. Nonetheless, the substantial increase in efficiency of IP.LSH.DBSCAN over regular DBSCAN enables rapid parameter exploration and makes the analysis pipeline practical for daily operations, even as the data scale increases.

Domain-Specific Clustering Results

After the initial exploration phase, we utilized our toolchain's feature engineering component to extract domain-specific characteristics from the consumption data. The feature extraction component is designed to be configurable and extensible, allowing domain experts to define and refine features relevant to their specific analytical questions without modifying the underlying system.

To demonstrate the toolchain's capabilities, we implemented a clustering approach using key features relevant to grid management: peak-to-average ratio, number of significant peaks, peak width, and base load. Analysis revealed six distinct consumption patterns, visualized in the Uniform Manifold Approximation and Projection (UMAP) [9] Figure 2. The clusters represent varying consumption behaviors differentiated by time of day (morning peaks in clusters 0 and 4, midday/afternoon peaks in clusters 1 and 5, evening peaks in clusters 2 and 3) and peak structure (single versus multiple significant peaks).

Clustering provides valuable insights for grid management applications, enabling, for example, time-of-use pricing optimization informed by consumption behavior. The clear separation between clusters in the UMAP visualization confirms the effectiveness of the chosen features in discriminating between distinct consumption behaviors. These findings demonstrate how feature engineering designed for energy consumption analysis can support more personalized approaches to grid management and customer segmentation.

6 Conclusions and Future Work

This paper presents a scalable multi-stage analysis pipeline for extracting meaningful insight from high-dimensional electricity consumption data. Our approach combines the computational efficiency of approximate clustering with the interpretability advantages of feature engineering, providing a practical toolchain for grid operators and analysts.

Initial applications at Göteborg Energi demonstrate improvements in computational performance and insight quality compared to traditional methods. The pipeline has successfully identified consumer behaviors relevant to grid management while maintaining computational feasibility for large datasets. In ongoing and continued work, we target to:

(1) Expand feature engineering capabilities to address additional behavioral questions
(2) Incorporate temporal tracking to monitor behavior evolution over extended periods
(3) Evaluate the generalizability of our approach to other domains with high-dimensional time series data
(4) Generalize the system itself to allow for users to define an arbitrary number of stages

Our results suggest that implementing a toolchain that combines efficient exploration with targeted feature engineering offers a promising direction for practical analysis of high-dimensional time series data in environments that require efficient resource use.

References

[1] Charu C. Aggarwal, Philip S. Yu, Jiawei Han, and Jianyong Wang. 2003. A Framework for Clustering Evolving Data Streams. In *Proceedings 2003 VLDB Conference*, Johann-Christoph Freytag, Peter Lockemann, Serge Abiteboul, Michael Carey, Patricia Selinger, and Andreas Heuer (Eds.). Morgan Kaufmann, San Francisco, 81–92. doi:10.1016/B978-012722442-8/50016-1

[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering–a decade review. *Information systems* 53 (2015), 16–38. doi:10.1016/j.is.2015.04.007

[3] Apache Software Foundation. 2024. *Apache Kafka: A Distributed Streaming Platform.* https://kafka.apache.org

[4] Ira Assent. 2012. Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery* 2, 4 (2012), 340–350. doi:10.1002/widm.1062

[5] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. 2015. Apache Flink: Stream and Batch Processing in a Single Engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 36, 4 (2015).

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231. doi:doi/10.5555/3001460.3001507

[7] Zhang Fu, Magnus Almgren, Olaf Landsiedel, and Marina Papatriantafilou. 2014. Online temporal-spatial analysis for detection of critical events in Cyber-Physical Systems. In *2014 IEEE International Conference on Big Data (Big Data)*. 129–134. doi:10.1109/BigData.2014.7004221

[8] Amir Keramatian, Vincenzo Gulisano, Marina Papatriantafilou, and Philippas Tsigas. 2022. IP.LSH.DBSCAN: Integrated Parallel Density-Based Clustering Through Locality-Sensitive Hashing. In *European Conference on Parallel Processing*. Springer, 268–284.

[9] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018). doi:10.48550/arXiv.1802.03426

[10] M. Muthamizh Selvam, R. Gnanadass, and N.P. Padhy. 2016. Initiatives and technical challenges in smart distribution grid. *Renewable and Sustainable Energy Reviews* 58 (2016), 911–917. doi:10.1016/j.rser.2015.12.257

[11] Statistiska Centralbyrån. 2025. *Swedish Standard Industrial Classification (SNI).* https://www.scb.se/en/documentation/classifications-and-standards/swedish-standard-industrial-classification-sni/

[12] Joris Van Rooij. 2020. Data stream processing meets the Advanced Metering Infrastructure: possibilities, challenges and applications.

[13] Michel Verleysen and Damien François. 2005. The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Computational Intelligence and Bioinspired Systems*, Joan Cabestany, Alberto Prieto, and Francisco Sandoval (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 758–770. doi:10.1007/11494669_93