XXVII International Symposium on Nuclear Electronics & Computing

Montenegro, Budva, Becici, 30 September - 4 October 2019

A study on performance assessment of essential clustering algorithms for the Interactive Visual Analysis toolkit **InVEx**

Mikhail TITOV, Maria GRIGORIEVA, Aleksandr ALEKSEEV, Nikita BELOV, Timofei GALKIN, Dmitry GRIN, Tatiana KORCHUGANOVA, Sergey ZHUMATIY

NEC'2019

Lomonosov Moscow State University

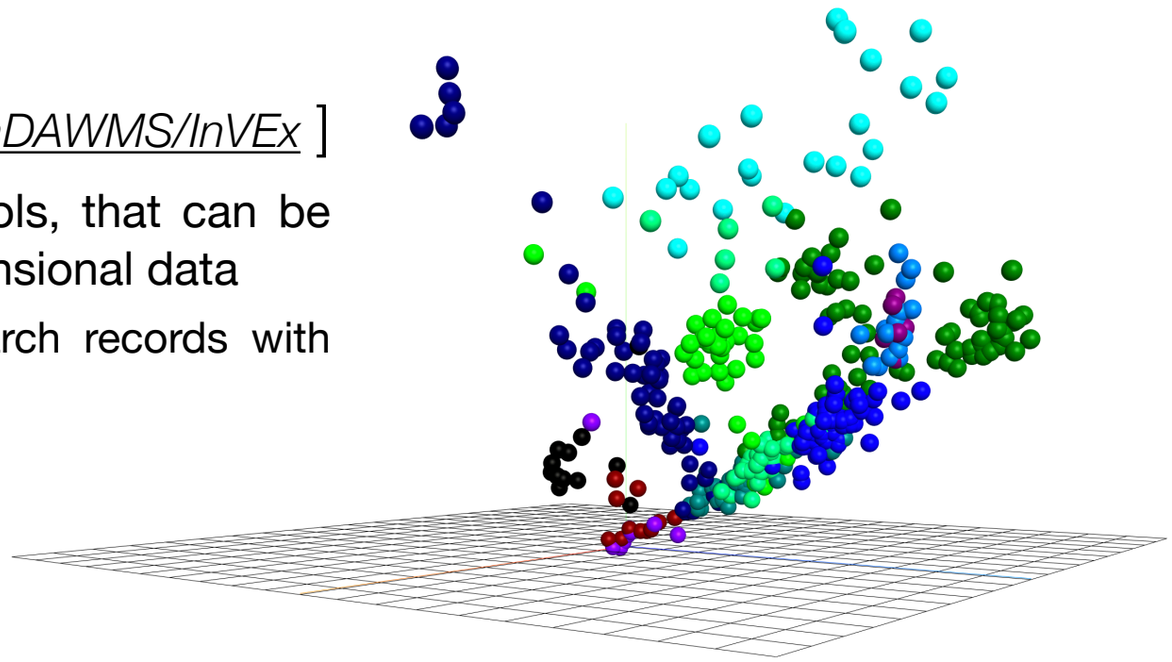
Introduction

InVEx overview

InVEx
Interactive Visual Explorer

Interactive Visual Explorer [*https://github.com/PanDAWMS/InVEx*]

- Provides advanced interactive data visualization tools, that can be applied to the analysis of large volumes of multidimensional data
  - Emphasis on interactive visual clustering (search records with non-trivial parameters and its possible reasons)



Framework

- **Django** (version 2) - Python-based web framework
- **Three.js** - cross-browser JavaScript library and API used to create and display animated 3D computer graphics in a web browser (uses WebGL)

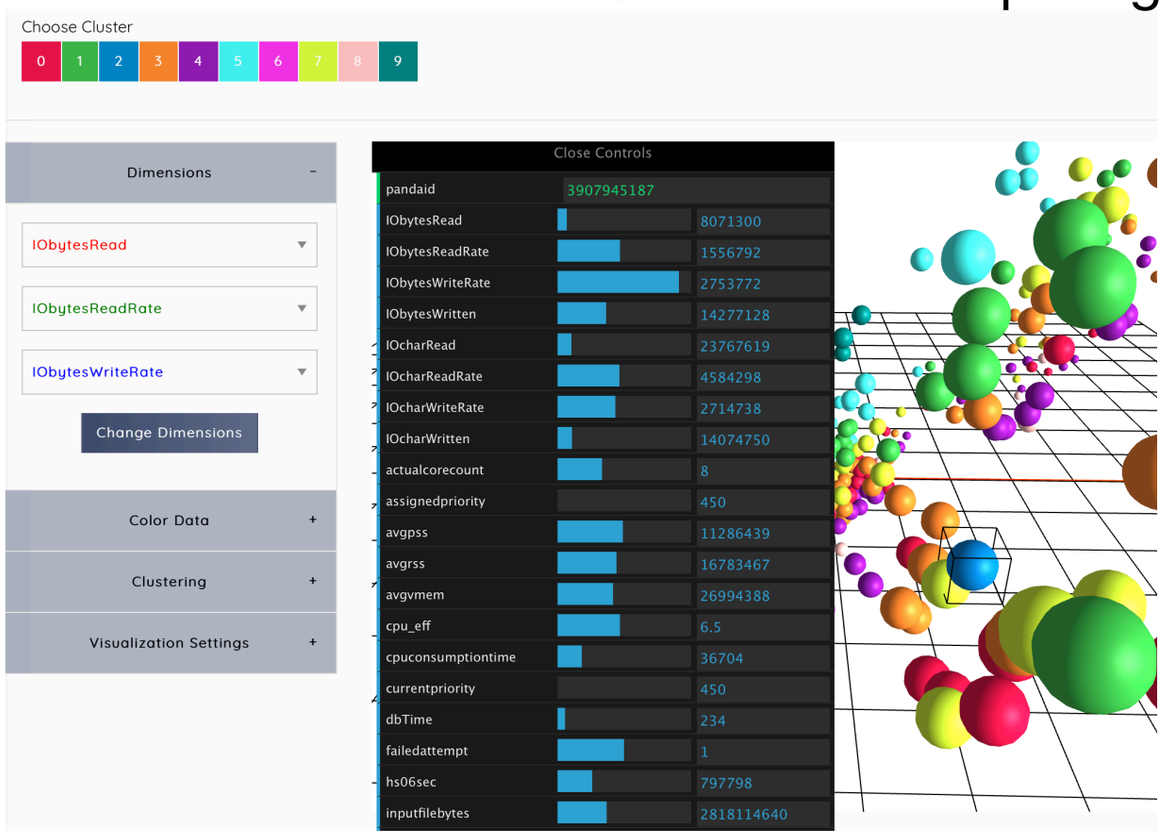
Analysis libraries

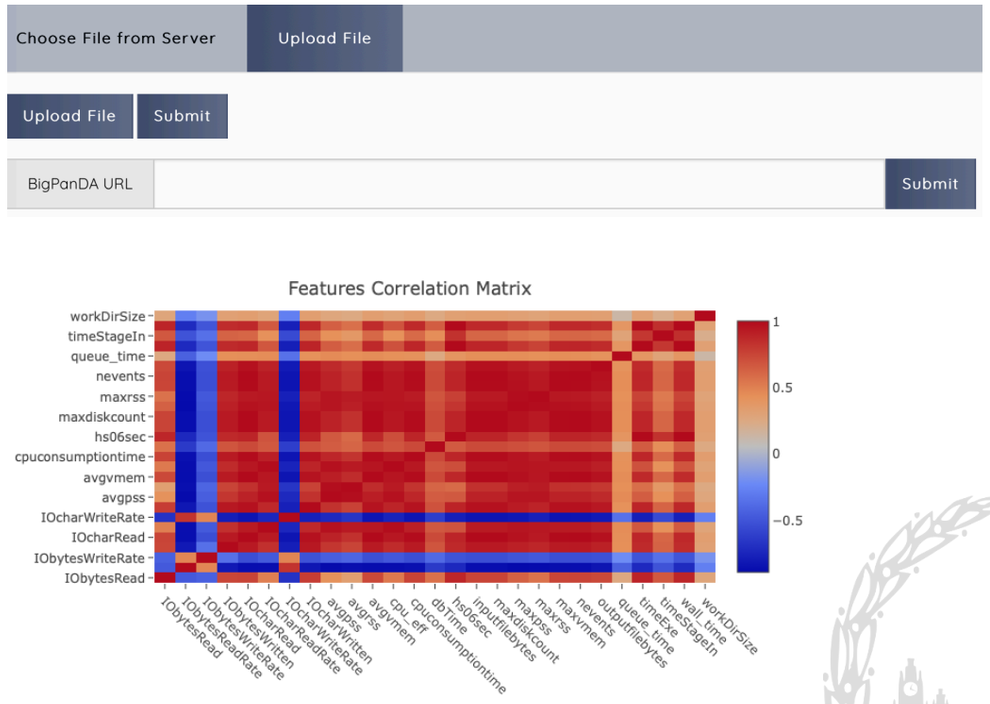
- **Pandas** - data manipulation and analysis
- **Scikit-learn** - machine learning library in Python
- **Extra libraries** *(for further quality improvements)*
  - **Prince** - factor analysis (aims to find independent latent variables)
  - **Hdbscan** - HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise
  - **Intel Data Analytics Acceleration Library (Intel DAAL)** - optimized algorithmic building blocks for data analysis stages

InVEx | First steps

ATLAS computing metadata has become the research ground

- Large amount of ATLAS ProdSys2/PanDA metadata provides a means to test and prove the efficiency of the applied technologies and methods
  - Initial integration with PanDA includes the direct access to information about computing jobs from BigPanDA Monitor system



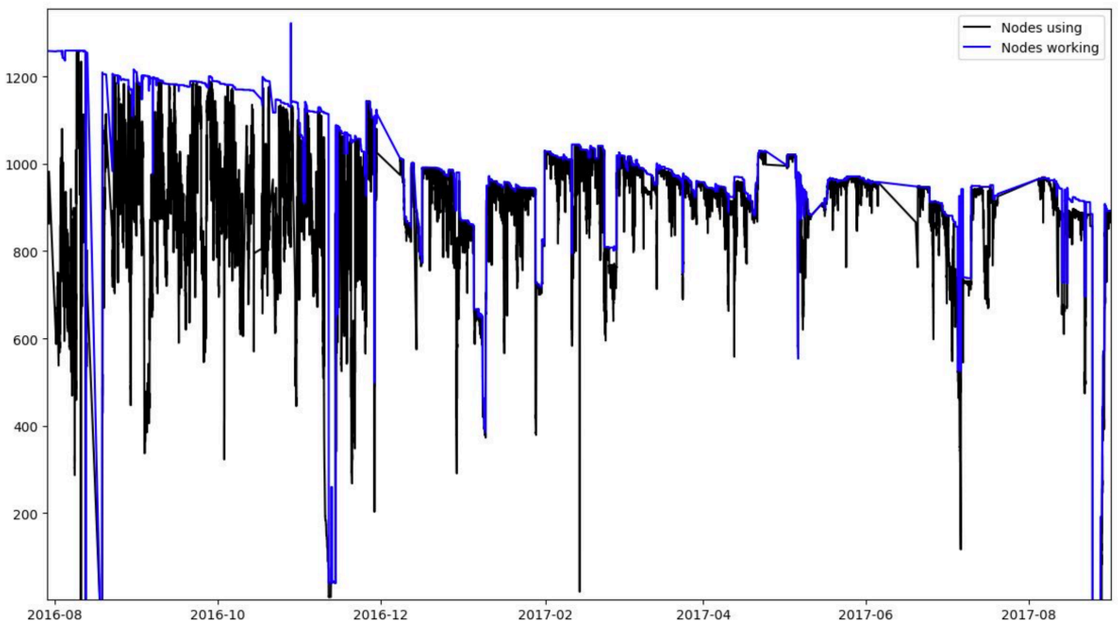


SC Lomonosov overview

Lomonosov-2 Supercomputer is designed by T-Platforms and installed at the Lomonosov Moscow State University (rank #93 in the list TOP500*)

- Intel Xeon/FDR InfiniBand cluster, accelerated with NVIDIA Tesla K40s and Tesla P100 GPUs
  - 1696 nodes (Intel Haswell-EP E5-2697v3, 2.6 GHz, 14 cores and Intel Xeon Gold 6126 2.60GHz, 12 cores) with 64/96 GB of memory per node
  - Peak performance 4.946 PFLOPS

** TOP500 - Lomonosov 2 - https://www.top500.org/system/178444*



*Lomonosov-2 supercomputer CPU time utilization (from August 2016 to August 2017)***

*** S Leonenkov et al. "Supercomputer Efficiency: Complex Approach Inspired by Lomonosov-2 History Evaluation" (2018)*

Clustering

Clustering algorithms

Clustering - grouping similar objects into unlabeled groups (unsupervised learning)

- *Data points within a particular cluster has to be very similar to the other data points (in that cluster), i.e., the within-cluster homogeneity has to be very high but on the other hand, the objects of a particular cluster have to be as dissimilar as possible to the objects present in other cluster(s)*

  - **Partitioning-based methods** - distance-based methods
  - **Density-based methods** - a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density
  - **Hierarchical methods** - split the data points into levels / hierarchies based on their similarities

Clustering validation

- **External measures**

  Comparison of the identified clusters to an external reference

  - **Adjusted Rand Index** - a function that measures the similarity of the two assignments (*labels_true* - given the knowledge of the ground truth class assignments; *labels_pred* - clustering algorithm assignments), ignoring permutations and with chance normalization
  - **Fowlkes-Mallows score** (Fowlkes-Mallows index) - defined as the geometric mean of the pairwise precision and recall

- **Internal measures**

  Reflection of the compactness, the connectedness and the separation of the cluster partitions

  - **Silhouette coefficient** - measures how well an observation is clustered and estimates the average distance between clusters
  - **Calinski-Harabaz Index** (Variance Ratio Criterion) - a ratio between the within-cluster dispersion and the between-cluster dispersion
  - **Davies-Bouldin Index** - based on the approximately estimation of the distances between clusters and their dispersion to obtain a final value that represents the quality of the partition

Clustering for InVEx

- **Initial data grouping - Level-of-Detail definition**

  Reduce the amount of data presented to the user

  - Clustering algorithms: MiniBatchKMeans, KPrototypes
  - Group by nominal / ordinal parameters - usage of categorical features for grouping
  - Group by numerical continuous parameter - splitting continuous parameter into the defined number of ranges

- **Cluster analysis**

  In-depth study of implicit correlations between multidimensional data objects

  - KMeans, MiniBatchKMeans, KPrototypes, DBSCAN, etc.

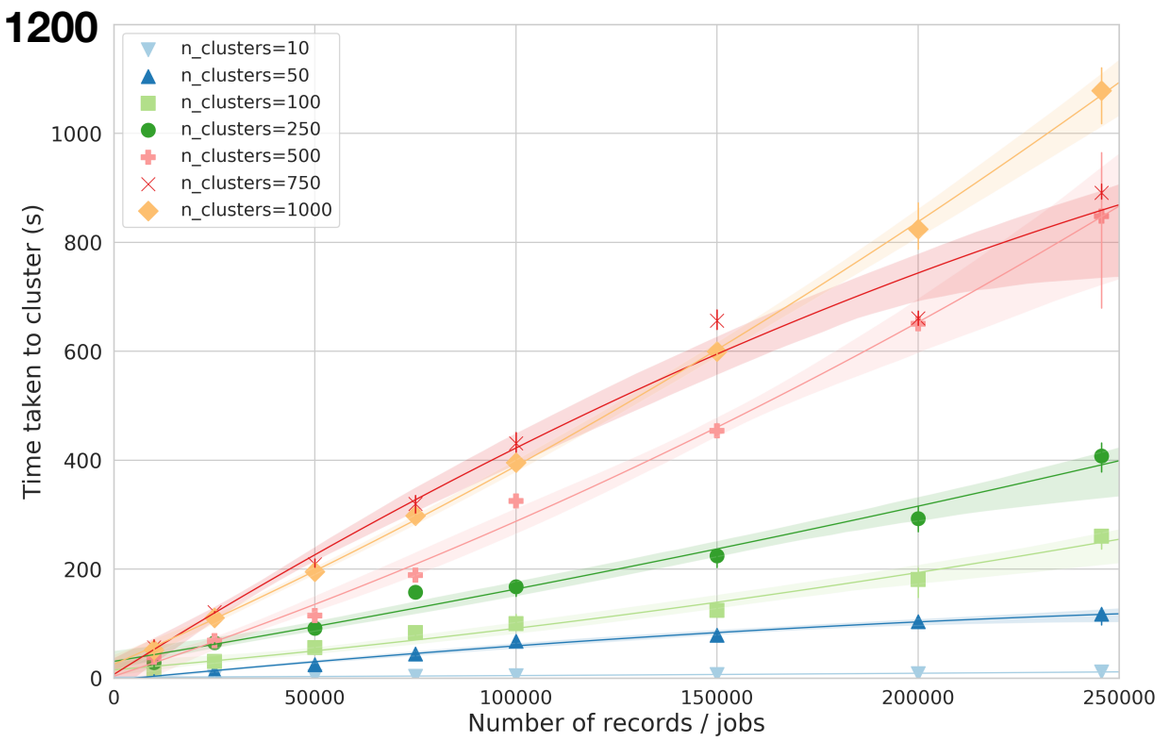
Experiments

Experimental data

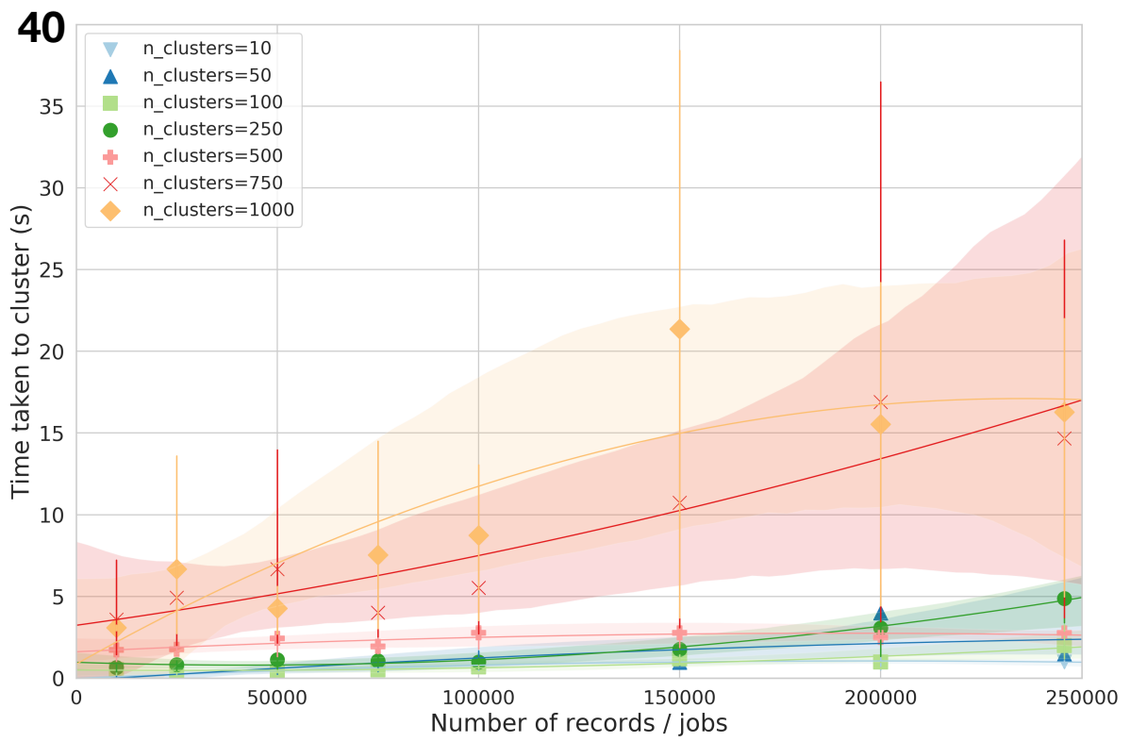
- **Data description**
  - Log data from SC Lomonosov for 300 days (from June 2018 to March 2019)
    - 245K records with **12** attributes (*user ID, execution time duration, number of allocated nodes, CPU load during the job execution per user, GPU load during the job execution per user, number of executed instructions per second, etc.*)
- **Data pre-processing techniques**
  - Dimensionality Reduction - mapping of the data to a lower-dimensional space
    - Linear techniques
      - Multiple correspondence analysis (MCA), Principal component analysis (PCA)
    - Non-linear techniques
      - t-Distributed Stochastic Neighbor Embedding (t-SNE)
      - Uniform Manifold Approximation and Projection (UMAP) - visualization similarly to t-SNE, and also general non-linear dimension reduction
- **Transformed data**
  - 11 attributes + MCA => 5 components + 1 attribute + PCA => 5 components

KMeans | Scikit-learn

KMeans



MiniBatchKMeans

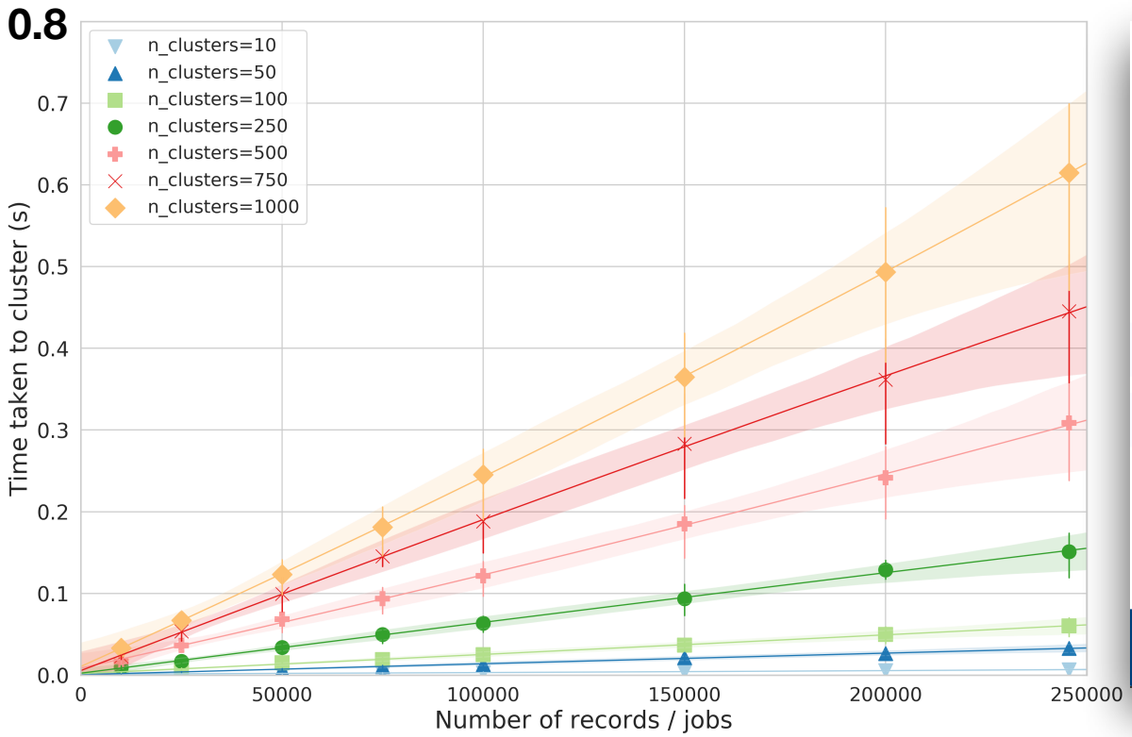


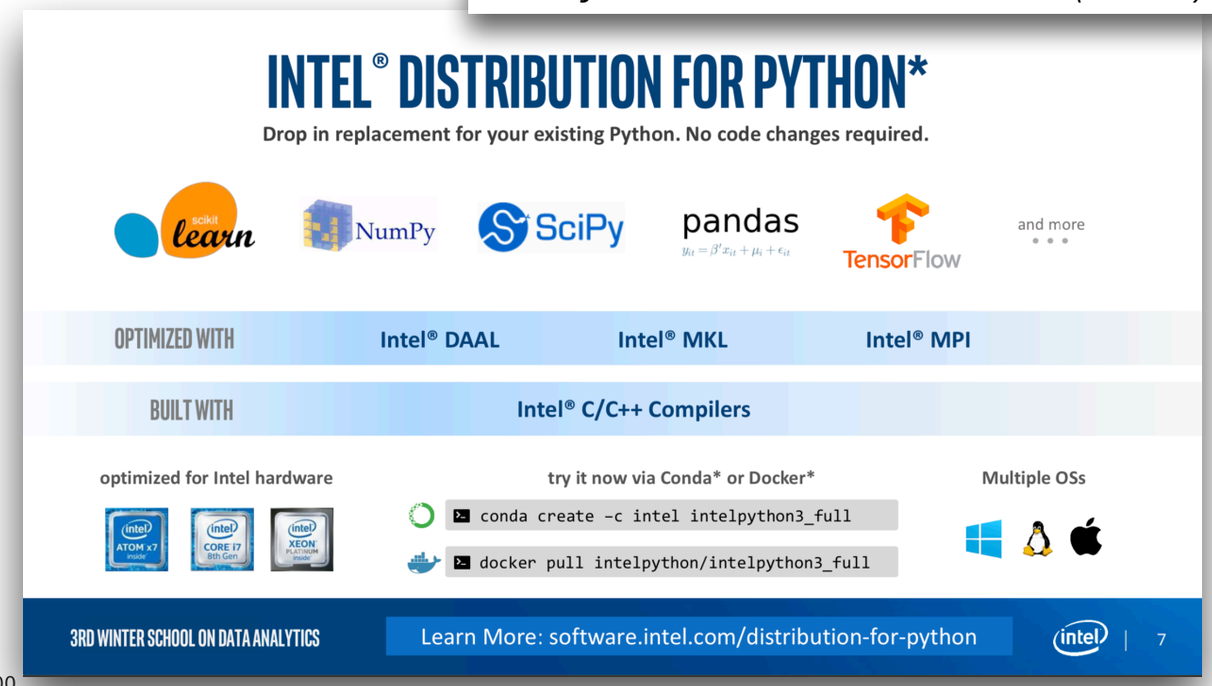
		KMeans	MiniBatchKMeans
silhouette_score	[**-1** : **1**]	0.61	0.58
calinski_harabaz_score	$+\infty$	92853.7	88390.1
davies_bouldin_score	[**0** : $+\infty$)	0.86	0.94

KMeans | Intel® DAAL

Intel® Data Analytics Acceleration Library

** R Israfilov, "Fast Data Analytics with Python and Intel DAAL" (2018)*

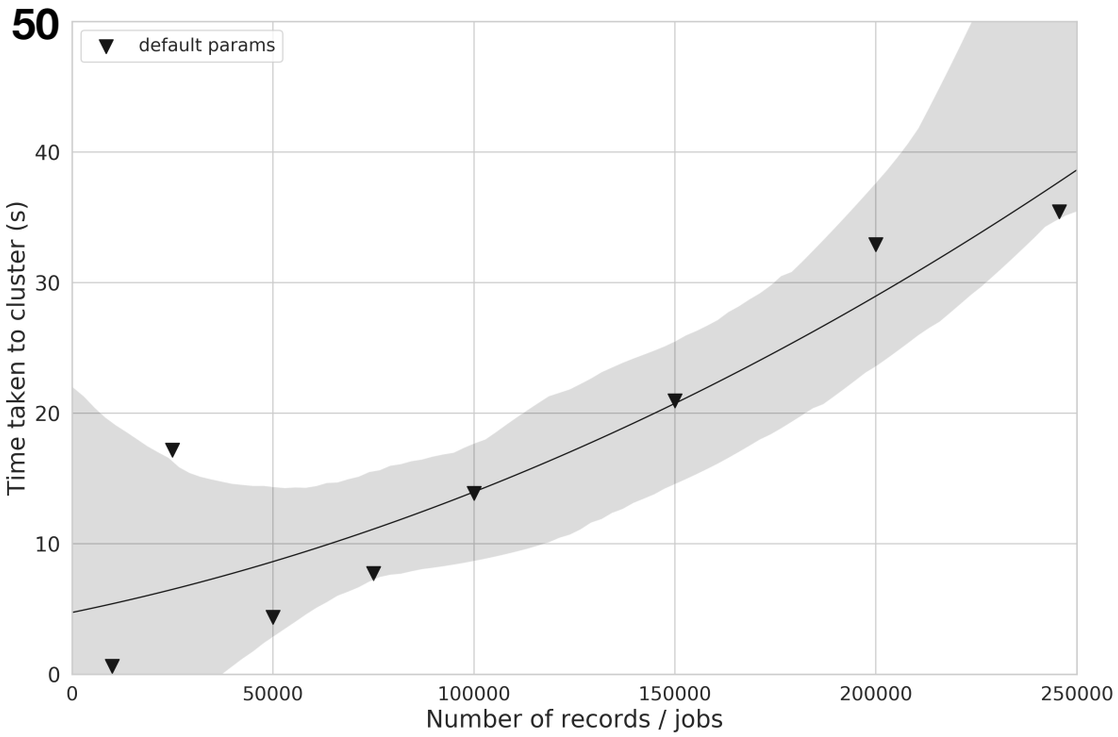




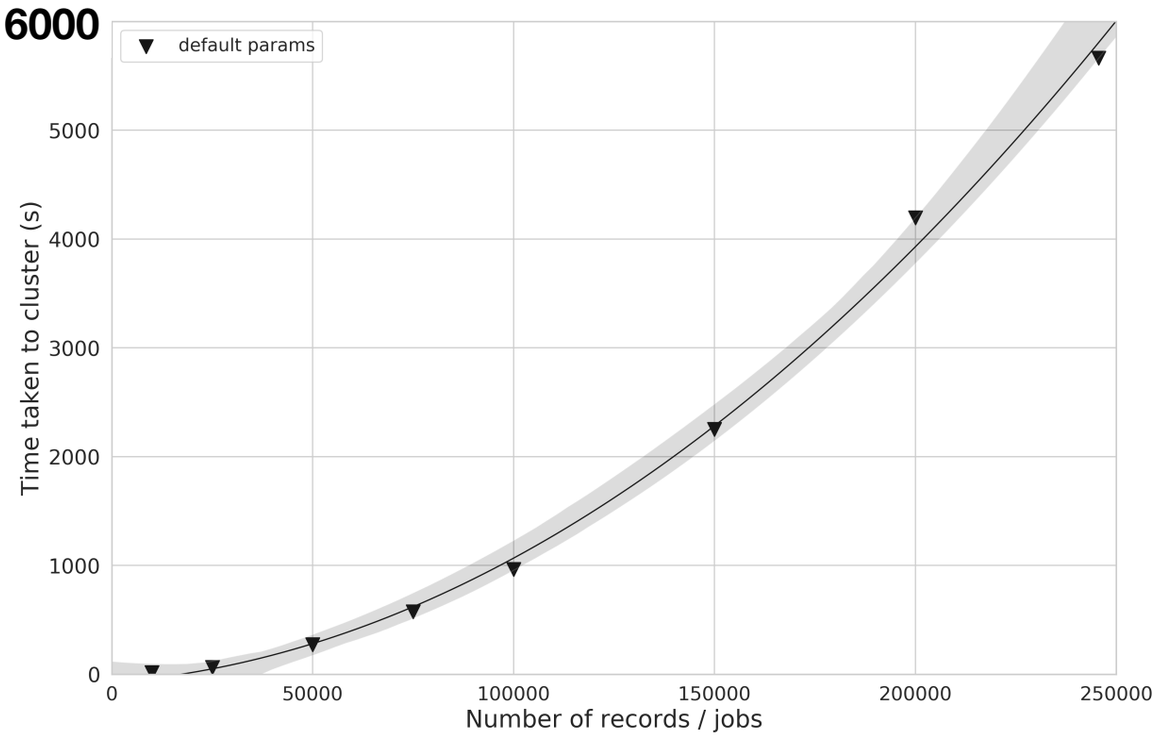
		KMeans
silhouette_score	[**-1** : **1**]	0.68
calinski_harabaz_score	$+\infty$	32503.6
davies_bouldin_score	[**0** : $+\infty$)	1.08

HDBSCAN & OPTICS

HDBSCAN



OPTICS



		HDBSCAN	OPTICS
silhouette_score	[**-1** : **1**]	0.58	0.59
calinski_harabaz_score	$+\infty$	211.5	73.4
davies_bouldin_score	[**0** : $+\infty$)	1.37	1.36

Pairs of algorithms

	adjusted_rand_score [**-1** : **1**]
(scikit-learn) KMeans vs. MiniBatchKMeans	0.55
scikit-learn / KMeans vs. scipy / KMeans	0.77
scikit-learn / MiniBatchKMeans vs. scipy / KMeans	0.74
scikit-learn / KMeans vs. daal4py / KMeans	**0.95**
HDBSCAN vs. OPTICS	0.84
MiniBatchKMeans vs. OPTICS&HDBSCAN	0.25

Conclusion

Summary

- InVEx represents a visual analytics approach aimed at cluster analysis and in-depth study of implicit correlations between multidimensional data objects
    - Originally designed to enhance the analysis of computing metadata of the ATLAS experiment at LHC for operational needs, but also provides the same capabilities for other domains to analyze large amounts of multidimensional data
- Benchmark tests assess the relative performance between chosen clustering algorithms and corresponding metrics, and the quality of produced clusters
    - Obtained results will be used as guidelines in assisting users in a process of visual analysis using InVEx

Acknowledgements

- Many thanks to the whole InVEx team and colleagues from the Research Computing Center of Moscow State University (RCC MSU) for providing the data for our analysis and for their continued support

- This work was funded by the Russian Science Foundation grant No.18-71-10003