A New Visual Analytics Toolkit for ATLAS Computing Metadata

M A Grigorieva[1], A A Alekseev[2], T P Galkin[3], A A Klimentov[4], T A Korchuganova[2], I E Milman, S V Padolski[4], V V Pilyugin[3] and M A Titov[1] on behalf of the ATLAS Collaboration

[1] Lomonosov Moscow State University, Moscow, Russia
[2] National Research Tomsk Polytechnic University, Tomsk, Russia
[3] National Research Nuclear University "MEPhI", Moscow, Russia
[4] Brookhaven National Laboratory, Upton, NY, USA

E-mail: maria.grigorieva@cern.ch, mikhail.titov@cern.ch

Abstract. The ATLAS experiment at the Large Hadron Collider has a complex heterogeneous distributed computing infrastructure, which is used to process and analyse exabytes of data. Metadata are collected and stored at all stages of data processing and physics analysis. All metadata could be divided into operational metadata to be used for the quasi on-line monitoring, and archival to study the behaviour of corresponding systems over a given period of time (i.e. long-term data analysis). Ensuring the stability and efficiency of complex and large-scale systems, such as those in the ATLAS Computing, requires sophisticated monitoring tools, and the long-term monitoring data analysis becomes as important as the monitoring itself. Archival metadata, which contains a lot of metrics (hardware and software environment descriptions, network states, application parameters, errors) accumulated for more than a decade, can be successfully processed by various machine learning (ML) algorithms for classification, clustering and dimensionality reduction. However, the ML data analysis, despite the massive use, is not without shortcomings: the underlying algorithms are usually treated as "black boxes", as there are no effective techniques for understanding their internal mechanisms. As a result, the data analysis suffers from the lack of human supervision. Moreover, sometimes the conclusions made by algorithms may not be making sense with regard to the real data model. In this work we will demonstrate how the interactive data visualization can be applied to extend the routine ML data analysis methods. Visualization allows an active use of human spatial thinking to identify new tendencies and patterns found in the collected data, avoiding the necessity of struggling with the instrumental analytics tools. The architecture and the corresponding prototype of Interactive Visual Explorer (InVEx) - visual analytics toolkit for the multidimensional data analysis of ATLAS computing metadata will be presented. The web-application part of the prototype provides an interactive visual clusterization of ATLAS computing jobs, search for computing jobs non-trivial behaviour and its possible reasons.

ATL-SOFT-PROC-2019-005
29 May 2019

1. Introduction

In the era of Big Data and mega-science projects, the data analysis becomes one of the most actively developing and important discipline. The more complex and diverse the data becomes, the more complicated the process of data analysis becomes. Statistics, machine and deep learning, and other data analysis techniques are widely used to classify, cluster and/or categorize data, to search for anomalies, predict objects behaviour in certain circumstances, to search correlations among data and discover its hidden peculiarities. But when the analyst deals with the huge amount of multidimensional

data, it is not an easy task to make sense of such data. Moreover, data analysis methods, such as machine learning (ML), which have been developed over the last decades, are usually treated as "black boxes", as there are limited effective techniques for understanding their internal mechanisms.

The human brain is barely able to process large and complex data, including understanding hidden processes in ML algorithms efficiently without auxiliary graphical tools, like plots, graphs, diagrams or charts. However, static visual images may hide important characteristics of data. To investigate data objects more efficiently, analysts need an interaction with graphical objects.

The approach of using visual interactive interfaces to facilitate analytical reasoning is called visual analytics [1]. This approach allows analysts to combine human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems.

The ATLAS experiment [2] at the LHC is one of the most representative examples of data complexity and volumes. ATLAS computing infrastructure, used to process and analyse exabytes of data, has a complex and changeable hierarchy structure of tiers, clouds, virtual organizations, federations, sites and nodes. All computing resources are geographically distributed worldwide. Physics analysis tasks executed within this environment have a complex workflow, containing multiple stages, implementing different automatic operational workloads. Like any complicated system, ATLAS computing infrastructure utilizes sophisticated monitoring tools and long-term analytics methods to ensure the stability and efficiency of its functioning. And the visual analytics methods may significantly simplify a long and complicated process of data analysis. This paper describes the designed and prototyped visual analytics toolkit - InVEx (Interactive Visual Explorer) providing interactive visual tools for the data analysis supervision. The main objectives of the proposed toolkit are the following:

- Support the sense-making process of data analysis by interactive visual tools;
- Increase the domain experts involvement in the process of data analysis;
- Enhance the routine data analysis methods (statistics, machine learning) with the use of visual interaction with the initial data and with the underlying algorithms as well.

2. Initial use cases for Visual Analytics in ATLAS Computing

As ATLAS computing metadata is used to demonstrate the visual analytics approach, we highlighted some use cases where the proposed approach may be useful.

- The analysis of computing jobs executions at the workload management system PanDA [3].
    - The main objective of such analysis is searching for the characteristics of non-trivial executed jobs and detection of possible reasons of such behaviour. Clusterization methods can be used to categorize jobs by similar features to search for anomalies among these groups. And the visual interactive interfaces may facilitate the clustering supervision and interpretation.
- The analysis of computing sites performance and robustness.
    - Visual analytics approach will allow to observe correlations between jobs execution processes on computing sites and site efficiency metrics. These correlations may help to understand jobs execution patterns in certain circumstances and consider them in future for transfers optimization.
- Utilization of the interactive GUI to observe changes of the correlations among data parameters over time (dynamic visualization).

3. The development of InVEx for ATLAS Computing

3.1. Essential considerations

Since the developed toolkit is not limited to one field of the application, it should include easily updated, adjusted and supported software components, thus, to be based on open-source technologies. Its accessibility and distribution are other key points that lead to building it as a web application,

which can be prototyped quickly and can combine high-quality graphics with the functionality of a browser (web-based GUI that supports flexibility and a high degree of interactivity).

The current prototype implementation of InVEx is focused on the computing environment of the ATLAS experiment, therefore it includes modules for its integration with the internal ATLAS metadata storages and information systems. It is worth mentioning that some of these sources are built based on Elasticsearch engine (e.g. ATLAS archived data management and networking metadata [4], including corresponding perfSONAR [5] data, DKB [6]) and have the same approach for data collection, others provide corresponding APIs (e.g. AGIS [7], Rucio [8], etc.). The efficiency of the designed application is determined by providing fast and flexible search interfaces for ATLAS metadata storages (i.e. collected data samples) and the results of applied data analysis.

3.2. Technologies

The choice of technologies resulted in the following libraries, packages and frameworks: i) web framework - Python [9] based Django framework [10]; ii) 3D visualization - JavaScript library Three.js [11]; iii) 2D visualization - JavaScript libraries D3.js [12] and Plotly.js; iv) ML/clustering algorithms - Python library Scikit-learn [13].

3.3. General overview

The InVEx application client has been built upon the concept of thick client to ensure highly interactive and responsive GUI. All calculations, including data preparation, normalization and clusterization, are executed on the server side. The general workflow of data processing and visualization is presented in Figure 1.

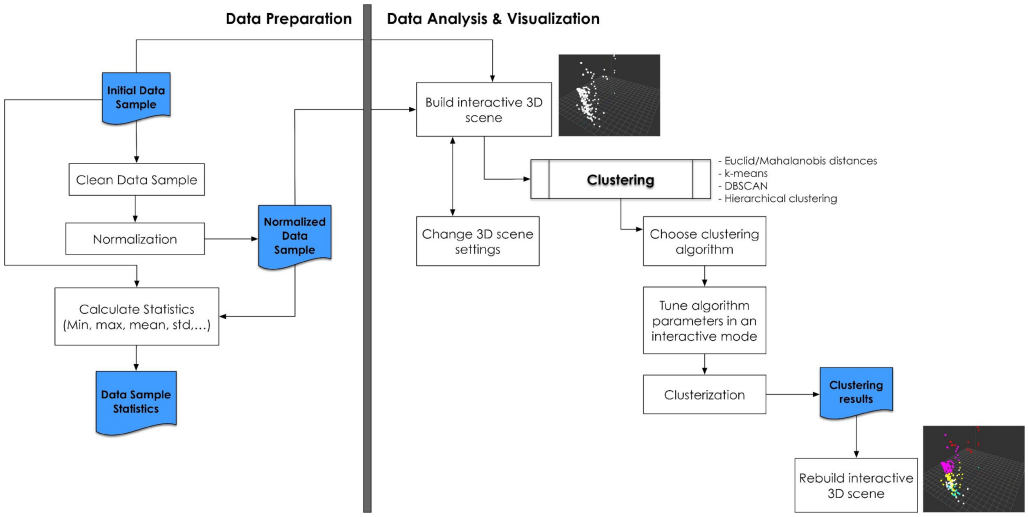

Figure 1. The workflow of the InVEx application.

The initial data sample represents raw data gathered for analysis, which is either uploaded as a local file or collected from the remote source (e.g. BigPanDA Monitor, Elasticsearch instance). The data processing goes through the following chain of transformations:

- Cleaning data sample from NaN values (i.e. a numeric data type value representing an undefined or unrepresentable value);
- Normalization;
- Calculating the descriptive statistics to summarize all features of the data sample.

Normalized data sample is used to build three-dimensional spatial scene, which contains a plane, three axes and spheres representing data objects. Thus, after data sample is prepared for the analysis, it is represented to the user as an interactive 3D scene. These spheres can be grouped, clusterized, picked

by a corresponding value (or range of values) of the defined object parameter, etc. The functionality of the InVEx application includes interactive handling of 3D projections for the analysis and analytical processes. User's actions (e.g. applied conditions, data (re)grouping and clusterization) can be saved as operations history to provide means to compare different approaches of data analysis and their results.

3.4. The prototype of the InVEx application for ATLAS Computing

InVEx is represented as a web application, that provides a set of instruments to conduct an analysis over defined data sample. The GUI contains the following panels:

- *Main Panel* - the visualization canvas with 3D visualization;
- *Auxiliary Data Panel* - data sample views in tabular form;
- *Control Panel* includes four control boxes:
    - *Projections:* provides means to choose the dimensions for 3D projection;
    - *Color Data:* grouping and highlighting data objects by specific parameters values;
    - *Clustering*: provides a set of clustering algorithms with corresponding setup parameters;
    - *Visualization Settings:* allows to manipulate the visual representation of analyzed data.

The current implementation of the InVEx prototype uses data samples of PanDA jobs that were collected from the University of Chicago Elasticsearch instance or obtained directly from the BigPanDA Monitor. After the initial data are loaded and submitted, the following actions are provided for the user to work with displayed 3D objects: spatial scene manipulation (rotating, scaling, changing 3D projections), individual object handling (representation of the full parameters set, displacement), handling group of objects (highlighting by the defined parameter of a certain value, statistics overview). The visualization process is fully interactive (the visual data representation is rebuilt after new parameters and conditions are applied). Initially, all spheres on a 3D scene are colored with the same color. The highlighting of spheres with different colors is applied for clusterization and grouping of data objects and helps to investigate and compare visually the characteristics of obtained clusters and groups. The process of clusterization is one of the key features of InVEx, currently it is performed using K-Means and DBSCAN algorithms. In the near future, the set of algorithms will be extended. Figure 2 shows the clustered visual data representation in 3D projection. Colored boxes in the bottom are used to get statistics overview for corresponding clusters.

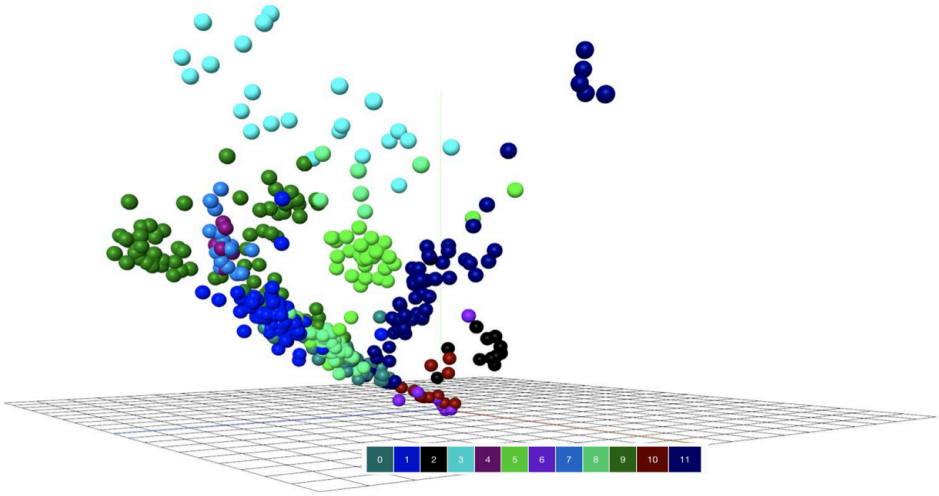

Figure 2. Data objects clusterization of PanDA jobs (dimensions: X axis - red - IObytesRead, Y axis - green - IObytesReadRate, Z axis - blue - IObytesWriteRate; number of clusters is 12).

The interactivity of the 3D scene expands the analysis capabilities. As an example, let's consider a scenario where a single sphere is located far from all others in one of the projections. Changing the position of a such sphere (by clicking on it and moving it with the mouse) closer to clusters will reveal changes in corresponding feature values and will help to estimate its interrelation with others. The object feature values could be changed manually as well (so-called direct interaction). The recalculation of clusters (i.e. re-clusterization) will put this re-located sphere into another cluster providing further understanding of features contribution and of the object nature in general.

4. Conclusion

The core idea of the designed visual analytics toolkit and its approach is to provide comprehensive information about multidimensional data and building knowledge about analyzed data using not just quantitative metrics (e.g. statistics representation), but also a human spatial thinking for a deeper understanding of analyzed objects interconnections. Further development of InVEx involves the addition of more clusterization methods, implementation of the storage system for the clusterization results, designing new visualization methods suitable for different use cases, and extending the scope of application beyond the ATLAS experiment.

Acknowledgments

We gratefully acknowledge the financial support from the Russian Science Foundation (grant No.18-71-10003).

References

[1] Thomas J J and Cook K A 2005 *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (Los Alamitos, CA: IEEE Press)

[2] The ATLAS Collaboration 2008 *JINST* **3** S08003

[3] Barreiro F H *et al.* 2016 *EPJ Web Conf.* **108** 01001

[4] Elasticsearch (a distributed, RESTful search and analytics engine) instance at UChicago, http://atlas-kibana.mwt2.org [accessed 2019-04-17]

[5] perfSONAR (Performance focused Service Oriented Network monitoring ARchitecture), an open source toolkit for running network measurements across multiple domains. Available from https://github.com/perfsonar [accessed 2019-04-17]

[6] Aulov V, Golosova M, Grigorieva M, Klimentov A, Padolski S, Wenaus T 2018 *J. Phys.: Conf. Ser.* **1085** 032013

[7] Anisenkov A, Di Girolamo A, Alandes M, Karavakis E 2015 *J. Phys.: Conf. Ser.* **664** 062001

[8] Barisits M *et al.* 2019 Rucio - Scientific Data Management *Preprint* arXiv:1902.09857

[9] Python project, "Python" [software], version 3.7.3, 2019. Available from https://www.python.org/downloads/release/python-373/ [accessed 2019-04-17]

[10] Django project, "Django" [software], version 2.1.8, 2019. Available from https://docs.djangoproject.com/en/2.1/releases/2.1.8/ [accessed 2019-04-17]

[11] Three.js project, "three.js" [software], version r103, 2019. Available from https://github.com/mrdoob/three.js/releases/tag/r103 [accessed 2019-04-17]

[12] Bostock M, Davies J, Heer J, Ogievetsky V and community, "D3.js" [software], version 5.9.2, 2019. Available from https://github.com/d3/d3/tree/v5.9.2 [accessed 2019-04-17]

[13] Scikit-learn project, "scikit-learn" [software], version 0.20.2, 2018. Available from https://github.com/scikit-learn/scikit-learn/tree/0.20.2 [accessed 2019-04-17]