Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Guest Editors)

Towards an Adaptive Framework for Real-Time Visualization of Streaming Big Data

Amin M. Khan[1,2], Daniel Gonçalves[2] and Duarte C. Leão[1]

[1]Pentaho Corporation. Lisbon, Portugal
[2]INESC-ID Lisboa. Instituto Superior Técnico, Universidade de Lisboa. Lisbon, Portugal

Abstract

Big data poses new challenges and the need for flexible, interactive, and dynamic visualization techniques. Existing approaches, especially in enterprise data visualization with static graphics or interactive dashboards, are limited at the scale of big data, given the volume and diversity of data to consider. Streaming data further compounds on the problem with the need for real-time analytics and visualizations. On the data acquisition and collection side of things, traditional business analytics platforms are being extended with support for technologies such as Apache Spark for improvement in performance. However, for real-time data visualization for streaming data, it is necessary to go beyond Apache Spark with in-memory processing and new data visualization idioms. We propose a framework for the dynamic visualization of real-time streaming big data, resilient to both its volume and rate of change. Some of the different directions we explore include: (a) the efficient processing and consumption of streaming data; (b) the automated detection of relevant changes in the data stream, highlighting entities that merit a detailed analysis; (c) the choice of the best idioms to visualize big data, possibly leading to the development of new visualization idioms; (d) real-time visualization changes.

Categories and Subject Descriptors (according to ACM CCS): [Human-centered computing]: Visualization—Visualization systems and tools

1. Introduction

The visualization of streaming big data, given its scale, volume, variety and rate of change, poses new challenges and the need for flexible, interactive, and dynamic visualization techniques. The challenges raised by streaming start at the most fundamental level with the need for a system that can cope with data that keeps pouring in. Even if that channel is properly managed, more interesting, high-level challenges arise. The first has to do with the detection of points of interest in the data. Indeed, an ordinarily uninteresting stream can suddenly become relevant to a particular analysis, when the data therein changes in nature, due perhaps to some external event. An analyst must be made aware of these changes. This can be done statistically but can be enhanced by the use of domain-specific knowledge. Doing this in a general way is an open challenge.

Even if it is possible to detect interesting changes to the data stream, the question of how to visualize them remains. In a typical business intelligence scenario, the data set and relevant analyses are usually well known beforehand. Thus, it is possible to manually custom-tailor visualizations (usually under the guise of dashboards) that perfectly match the data and the questions analysts have. With streaming data, especially in rapidly evolving complex domains, different situations may arise where different visualization idioms may be more relevant. This immediately raises the question about how to (semi-)automatically choose which idioms to use at any moment in time, depending on the data and its properties at that moment. Also, data changes that require new idioms to be used, raise the challenge of how to transition between different states of the visualization while maintaining the context of the analysis, and while also highlighting what the new important elements is.

We propose a framework that allows for clear visualization of streaming big data, covering matters such as (a) the efficient processing and consumption of streaming data; (b) the automated detection of relevant changes in the data stream, highlighting entities that merit a detailed analysis; (c) the choice of the best idioms to visualize big data, possibly leading to the development of new visualization idioms; (d) real-time visualization changes, and the overall technology stack to support these visualizations.

To process the volume of big data in timely manner, one approach, MapD, resorted to massively parallel in-memory databases and GPUs for quick interactive visualizations [Mat17, Tal13]. Another approach is to generate approximate visualizations, sacrificing accuracy for speed, but preserving visual properties of the data features [KBP*15]. Exploratory browsing through dynamic prefetching of portion of data allows for interactive analysis of large

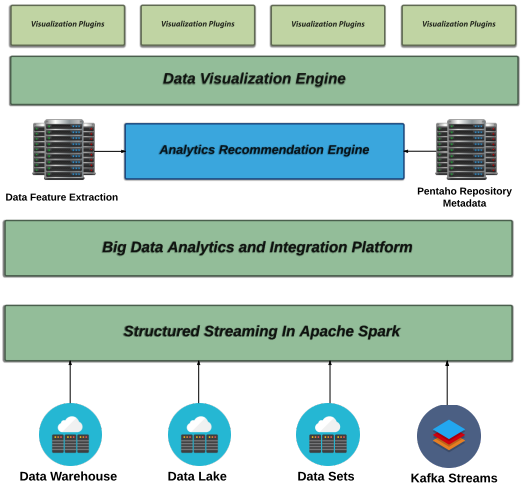

Figure 1: *Architecture*

datasets [BCS16]. For high-dimensional datasets, a visualization recommendation engine can assist in fast visual analysis [VRM*15]. Other works have explored related problems like preserving privacy when visually exploring data sets [HRM16], interactive data-driven visual query interfaces [BCD16], and visualization-oriented data aggregation and dimensionality reduction [JJHM14]. Alas, an integrated process, from data acquisition to visualization, is lacking.

2. Big Data Analytics and Visualization Framework

Figure 1 gives an overview of the big data integration and analytics framework we propose. It addresses the three key areas where innovation is required in order to support real-time visualization for streaming big data.

1. Data visualization focused data aggregation and dimensionality reduction at the data access layer, especially for streaming data (building on [JJHM14]).
2. Recommendation engine, employing machine learning, that builds on features in data sets (building on [VRM*15]), that not only helps with feature selection, but also recommends the appropriate visual idioms for presenting data.
3. Introducing and enhancing visualization idioms tailored for big data, with an emphasis on the smooth transition between idioms to better display items, attributes and patterns that become relevant as the stream's content changes.

The basis for the framework is Pentaho's data integration and business analytics platform [CBvD10] (although another similar platform could be used). Like other business analytics software for traditional data warehousing, Pentaho is not primarily designed for the real-time analytics use case. Therefore it needs to be enhanced and extended. To that end, Structured Streaming in Apache Spark can serve as a unified access layer for different kind of data sets and streams, with the help of approaches for the visualization of streaming data, such as Apache Kafka and Akka, accessible through Apache Spark.

The data integration and analytics layer connects to Apache Spark, and makes available the underlying data sets and streams to *Analytics Recommendation Engine*. The recommendation engine considers four different types of inputs to assist with big data visualization:

1. Contents, metadata and features from data sets
2. Advanced feature and metrics extraction engine, that employs machine learning for extracting useful feature sets from data
3. Metadata information gleaned from the data workflows and transformations created by the users in Pentaho platform
4. Domain-dependent knowledge

The recommendation engine should be able to collect all of the aforementioned data and not only make it available to the visualization engine but also identify relevant aspects of the data stream that become relevant. Changes in nature of the data, attributes that start to correlate, cyclic occurrences and reaching certain thresholds are mere examples of the types of situations the recommendation engine can detect. These may lead to the automatic computation of derived measures. Whenever these situations are detected, they as well will be made available to the visualization engine, that will then highlight the new information.

3. Big Data Visualization Engine

Existing visualization idioms can be enhanced to cope with large numbers of items/attributes by resorting to (a) Visual representations that while fairly limited in their encoding of information (ultimately, 1 px per datum) allow many to be displayed; (b) Abstract overview, coupled with "detailed detail" (drilling down, etc.) but in meaningful (and automated) ways; (c) Attribute and item reductions through statistical methods, clustering, etc.

The use of these techniques is, however, to a great extent, domain-specific. We propose that the (semi-) automatically adaptation to the data to be represented, accomplished with the help of the recommendation engine, is a better approach. Even more so, it will be a way to cope with the mutable nature of streaming data. Certain events or changes in the data may give rise to new patterns and particular aspects that must be made apparent in the visualization. This will require the use of specific visualization idioms and different encodings for the data. This usually implies the need for re-encoding (changing the visual idiom used) and/or reconfiguring (resorting, reordering, realigning, etc.) the visualizations. While some work has been done in these areas, how to do it automatically based on the nature of the data is an open issue. The main question to be answered is how to transition between a set of visualizations and a new one, better adapted to the data now in the stream, in such a way as not to confuse the analyst and still keep in context his analyses. This can pass for gradual changes passing through intermediate idioms, the careful use of animation, and compromises, where certain sub-optimal idioms are used since better ones would be disruptive of the analyst's work.

4. Conclusion

We have highlighted the need for a better visualization system tailored for streaming big data and proposed a framework that, while still in a design phase, can properly address that reality.

References

[BCD16] BHOWMICK S. S., CHOI B., DYRESON C.: Data-driven Visual Graph Query Interface Construction and Maintenance: Challenges and Opportunities. *Proceedings of the VLDB Endowment 9*, 12 (Aug. 2016), 984–992. 2

[BCS16] BATTLE L., CHANG R., STONEBRAKER M.: Dynamic Prefetching of Data Tiles for Interactive Visualization. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD)* (San Francisco, USA, June 2016), ACM Press, pp. 1363–1375. 2

[CBvD10] CASTERS M. R., BOUMAN R., VAN. DONGEN J.: *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley, 2010. 2

[HRM16] HE X., RAVAL N., MACHANAVAJJHALA A.: A Demonstration of VisDPT: Visual Exploration of Differentially Private Trajectories. *Proceedings of the VLDB Endowment 9*, 13 (Sept. 2016), 1489–1492. 2

[JJHM14] JUGEL U., JERZAK Z., HACKENBROICH G., MARKL V.: M4: A Visualization-Oriented Time Series Data Aggregation. *Proceedings of the VLDB Endowment 7*, 10 (June 2014), 797–808. 2

[KBP*15] KIM A., BLAIS E., PARAMESWARAN A., INDYK P., MADDEN S., RUBINFELD R.: Rapid Sampling for Visualizations with Ordering Guarantees. *Proceedings of the VLDB Endowment 8*, 5 (Jan. 2015), 521–532. 1

[Mat17] MATHESON R.: Split-Second Data Mapping. *MIT News* (Jan. 2017). URL: http://news.mit.edu/2017/startup-mapd-fast-big-data-mapping-0111. 1

[Tal13] TALBOT D.: Graphics Chips Help Process Big Data Sets in Milliseconds. *MIT Technology Review* (Oct. 2013). URL: https://www.technologyreview.com/s/520021/graphics-chips-help-process-big-data-sets-in-milliseconds/. 1

[VRM*15] VARTAK M., RAHMAN S., MADDEN S., PARAMESWARAN A., POLYZOTIS N.: SeeDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *Proceedings of the VLDB Endowment 8*, 13 (Sept. 2015), 2182–2193. 2