

# Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis

Pedro Casas (1), Alessandro D’Alconzo (1), Tanja Zseby (2), Marco Mellia (3)

(1) AIT Vienna, (2) TU Wien, (3) Politecnico di Torino

(1) name.surname@ait.ac.at, (2) tanja.zseby@tuwien.ac.at, (3) mellia@tlc.polito.it

## ABSTRACT

The complexity of the Internet has dramatically increased in the last few years, making it more important and challenging to design scalable Network Traffic Monitoring and Analysis (NTMA) applications and tools. Critical NTMA applications such as the detection of anomalies, network attacks and intrusions, require fast mechanisms for online analysis of thousands of events per second, as well as efficient techniques for offline analysis of massive historical data. We are witnessing a major development in Big Data Analysis Frameworks (BDAFs), but the application of BDAFs and scalable analysis techniques to the NTMA domain remains poorly understood and only in-house and difficult to benchmark solutions are conceived. In this position paper we describe the basis of the Big-DAMA research project, which aims at tackling this growing need by benchmarking and developing novel scalable techniques and frameworks capable to analyze both online network traffic data streams and offline massive traffic datasets.

## CCS Concepts

•Networks → Network measurement; Network monitoring; •Computing methodologies → Machine learning; Distributed computing methodologies;

## Keywords

Big Data; Network Traffic Monitoring and Analysis; Data Stream Processing; Machine Learning; Data Mining

The research leading to these results has been funded by the Vienna Science and Technology Fund (WWTF) through project ICT15-129, “BigDAMA”.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

LANCOMM, August 22-26, 2016, Florianopolis, Brazil

© 2016 ACM. ISBN 978-1-4503-4426-5/16/08... \$15.00

DOI: <http://dx.doi.org/10.1145/2940116.2940117>

## 1. WHY BIG-DAMA?

Network Traffic Monitoring and Analysis (NTMA) has taken a paramount role to understand the functioning of the Internet, especially to get a broader and clearer visibility of unexpected events. One of the major challenges faced by large-scale NTMA applications is the processing and analysis of large amounts of heterogeneous and fast network monitoring data. Network monitoring data usually comes in the form of high-speed streams, which need to be rapidly and continuously processed and analyzed. A variety of methodologies and tools have been devised to passively monitor network traffic, extracting large amounts of data from live networks. What is needed is a flexible data processing system able to analyze and extract useful insights from such rich and heterogeneous sources of data, offering the possibility to apply complex Machine Learning (ML) and Data Mining (DM) techniques. The introduction of Big Data processing led to a new era in the design and development of large-scale data processing systems. This new breed of tools and platforms are mostly dissimilar, have different requirements, and are conceived to be used in specific situations for specific needs. Each Big Data practitioner is forced to muddle through the wide range of options available, and NTMA is not an exception. A similar problem arises in the case of Big Data analytics through ML and DM based techniques. Despite the existence of ML libraries for Big Data Analysis Frameworks (BDAFs), there is a big gap to the application of such techniques for NTMA when considering fast online streams and massive offline datasets.

One of the questions a network operator poses himself when considering the NTMA domain is straightforward: if one wants to tackle NTMA applications with (near) real-time requirements in current massive traffic scenario, which is the best system for the task? If the main target is to perform complex data analytics on top of massive traffic, how should it be done? Which are the best ML/DM algorithms for doing so? The Big-DAMA project will accelerate NTMA practitioners’ and researchers’ understanding of the many new tools and techniques that have emerged for Big Data Analytics in recent years. Big-DAMA will particularly identify and test the most suitable BDAFs and available

Big Data Analytics implementations of ML and DM algorithms for tackling the problems of Anomaly Detection and Network Security in an increasingly complex network scenario. Big-DAMA has three main objectives: **Analytics**: conceive scalable online and offline ML- and DM-based techniques to monitor and characterize extremely fast and/or extremely large network traffic datasets; **Big-NTMA Frameworks**: conceive novel frameworks for Big Data Analytics tailored to Anomaly Detection and Network Security, evaluating and selecting the best BDAFs matching NTMA needs; and **Benchmarking**: conceive novel benchmarks for Big Data Analytics and BDAFs tailored to NTMA applications, particularly focusing on stream analysis algorithms and online processing tasks.

Big-DAMA aims at creating strong know-how in the principled application of Big Data analysis techniques and the usage of BDAFs in NTMA applications with online requirements. The starting point of Big-DAMA is DBStream [1], a Data Stream Warehouse we have recently developed and benchmarked against new big data analysis platforms such as Spark [8], showing very promising results in the field of NTMA [1]. DBStream has been running on a core ISP cellular network for more than 2 years, providing excellent performance and unprecedented network visibility.

A major challenge for any Big Data Analysis project is access to real big datasets. The developed techniques and selected frameworks will be experimentally evaluated with available Big Data traffic datasets, being these publicly available datasets or traffic traces captured by the partners of the project. Additional datasets come from external project partners such as CAIDA (<https://www.caida.org/data/>) and the MAWILab initiative (<http://mawi.wide.ad.jp/mawi/>).

## 2. AN OVERVIEW ON BIG NTMA

The introduction of Big Data processing led to a new era in the design and development of large-scale data processing systems [3]. A basic yet complete taxonomy of Big Data Analysis Frameworks includes traditional Database Management Systems (DBMS) and extended Data Stream Management Systems (DSMSs), noSQL systems (e.g., all the MapReduce-based systems), and Graph-oriented systems. While the majority of these systems target the offline analysis of static data, some proposals consider the problem of analyzing data coming in the form of online streams. DSMSs such as Gigascope [4] and Borealis [5] support continuous online processing, but they cannot run offline analytics over static data. The Data Stream Warehousing (DSW) paradigm provides the means to handle both types of online and offline processing requirements within a single system. DataCell and DataDepot are examples of this paradigm [2]. NoSQL systems such as MapReduce [6] have also rapidly evolved, supporting the analysis of unstructured data. Apache Hadoop [7] and Spark [8]

are very popular implementations of MapReduce systems. These are based on offline processing rather than stream processing. There has been some promising recent work on enabling real-time analytics in NoSQL systems, such as Spark Streaming [9], Indoop [10], Muppet [11] and SCALLA [12], but these remain unexploited in the NTMA domain. Besides these systems, there is a large range of alternatives, including Storm, Samza, Flink (NoSQL); Hawq, Hive, Greenplum (SQL-oriented); Giraph, GraphLab, Pregel (graph-oriented), as well as well known DBMSs commercial solutions such as Teradata, Dataupia, Vertica and Oracle Exadata (just to name a few of them).

The main challenge for algorithms for stream data processing are memory limitations and the resulting need to process data fast and in a single run. In [19], a Massive Online Analysis (MOA) software for classifying stream data is presented. The authors adapt several ML methods for stream data processing and provide performance criteria especially designed for evaluating stream data classification methods. Apache SAMOA (<https://samoa.incubator.apache.org>) is another streaming ML framework that provides algorithms for classification, clustering, and regression of streaming data. Clustering methods to support unsupervised learning for stream data are proposed for instance in [18] and [20]. Some methods are included in the MOA framework. Algorithms for stream clustering are also integrated in the statistics software R [21].

The application of BDAFs for NTMA tasks requires certain system capabilities: i) scalability: the framework must offer, possibly inexpensively, storage and processing capabilities to scale with huge amounts of data generated by in-network traffic monitors and collectors; ii) real-time processing: the system must be able to ingest and process data in real-time fashion; iii) historical data processing: the system must enable the analysis of historical data; iv) traffic data analysis tools: availability of embedded libraries or plugins specifically tailored to analyze traffic data.

In the following we present the main categories in which currently available data analysis technologies can be classified. For each of them, we highlight pros and cons, and explain why none of them fits for NTMA. Traditional SQL-like databases are inadequate for the continuous real-time analysis of data. As we mentioned before, DSWs have been introduced to extend traditional database systems with continuous data ingest and processing. These technologies leverage arbitrary SQL frameworks to perform rolling data analysis, i.e., they periodically import and process batches of data arriving at the system. In some cases, these technologies have been proven to be able to outperform – in terms of processing speed – Big Data technologies [1]. More recent solutions in this direction include ENTRADA [16], a Hadoop-based DSW for network traffic analysis, using off-the shelf Impala query engine and Parquet file format based on Google’s Dremel [17] to achieve

high performance, relying on columnar data storage. BDAFs based on the MapReduce paradigm have been recently started to be adopted for NTMA applications [13]. Considering the specific context of network monitoring, some solutions to adapt Hadoop to process traffic data have been proposed [14]. However, the main drawback of Big Data technologies in general is their inherent offline processing, which is not suitable for real-time traffic analysis, highly relevant in NTMA tasks. One of the few systems that leverage Hadoop for rolling traffic analysis is described in [15]. As said, there are also BDAFs for online data processing, but none of them has been yet applied to the NTMA domain.

### 3. SOME SCIENTIFIC CHALLENGES

There are several limitations in current state of the art related to the application of Big Data Analytics to NTMA applications. Firstly, Big Data Analytics' results on NTMA applications are seldom available, specially when considering online, stream based traffic analysis. This creates a major gap between the development of Big Data Analytics and BDAFs and the development of NTMA systems capable to analyze huge amounts of network traffic. In addition, while there is a vast number of BDAFs claiming different capabilities, the offer is so big and difficult to track that it is challenging to determine which one to choose for NTMA.

Secondly, considering the theory of Big Data Analytics applied to the NTMA domain, most of the proposed ML frameworks and libraries do not scale well in fast big data scenarios, as their main target is offline data analytics. In addition, while some supervised and unsupervised learning algorithms are already available for Big Data Analytics, we are at a very early stage development and there is big room for improvement. The most notable example is exploratory data analysis through clustering. Available algorithms are either too simple or too tailored to specific domains not related to traffic analysis. Clustering data streams is still an open problem, but very useful for unsupervised Anomaly Detection. Similar unsolved problems such as unsupervised feature selection become more challenging as well, due to scalability issues in the Big Data scenario. Also when considering supervised approaches, we do not have today much evidence on how supervised online learning approaches perform with big stream-based traffic.

There are also limitations in the benchmarking of different ML techniques running in BDAFs, because available benchmarks are very ad-hoc and tailored to specific types of systems.

### 4. PROSPECTIVE BENEFITS

The Big-DAMA project will advance the understating of the NTMA community on the aforementioned open problems. The outcomes of the Big-DAMA project have direct impact and application in the NTMA domain, including benefits for large network operators and

network monitoring technology vendors. The techniques developed within the span of the project as well as the application of BDAFs to online data analytic problems would also be highly beneficial to other domains where similar data analysis problems arise, including the online monitoring of M2M devices, the online extraction of knowledge from big data associated to smart cities scenarios, the processing of the avalanche of data generated by the upcoming Internet of Things with trillions of devices connected to the Internet, and many other application domains. Being the Big Data Analytics a fast-growing worldwide market, the development of analysis techniques, technologies, as well as strong know-how in the domain shall directly benefit the research NTMA community. To know more about the Big-DAMA project, visit <https://bigdama.ait.ac.at>.

### 5. REFERENCES

- [1] A. Baer et al., "Large-Scale Network Traffic Monitoring with DBStream, a System for Rolling Big Data Analysis," in *IEEE Big Data*, 2014.
- [2] L. Golab et al., "Stream Warehousing with DataDepot," in *SIGMOD*, 2009.
- [3] M. Stonebraker, "SQL Databases vs. noSQL Databases," in *Comm. of ACM*, vol. 53(4), 2010.
- [4] C. Cranor et al., "Gigascop: A Stream Database for Network Applications," in *SIGMOD*, 2003.
- [5] D. Abadi et al., "Aurora: A New Model and Architecture for Data Stream Management," in *VLDB J.*, 12(2), 2003.
- [6] J. Dean et al., "MapReduce: Simplified Data Processing on Large Clusters," in *Comm. of ACM*, 51(1), 2008.
- [7] T. White, "Hadoop: the Definitive Guide," *O'Reilly Media, Inc.*, ISBN:0596521979 9780596521974, 2009.
- [8] M. Zaharia et al., "Spark: Cluster Computing with Working Sets," in *HotCloud*, 2010.
- [9] M. Zaharia et al., "Discretized Streams: An Efficient and Fault-tolerant Model for Stream Processing on Large Clusters," in *HotCloud*, 2012.
- [10] P. Bhatotia et al., "Indoop: Mapreduce for Incremental Computations," in *ACM SoCC*, 2011.
- [11] W. Lam et al., "Muppet: Mapreduce-style processing of fast data," in *VLDB Endow.*, vol. 5(12), 2012.
- [12] B. Li et al., "Scalla: A Platform for Scalable one-pass Analytics using Mapreduce," in *ACM TODS* 37(4), 2012.
- [13] R. Fontugne et al., "Hashdoop: A MapReduce Framework for Network Anomaly Detection," in *Infocom Wks*, 2014.
- [14] Y. Lee et al., "Toward Scalable Internet Traffic Measurement and Analysis with Hadoop," in *SIGCOMM Comput. Commun. Rev.*, 43(1), 2012.
- [15] J. Liu et al., "Monitoring and Analyzing Big Traffic data of a Large-scale Cellular Network with Hadoop," in *IEEE Network*, 28(4), 2014.
- [16] M. Wullink et al., "ENTRADA: a High-Performance Network Traffic Data Streaming Warehouse," in *IEEE/IFIP NOMS*, 2016.
- [17] S. Melnik et al., "Dremel: Interactive Analysis of Web-scale Datasets," in *VLDB Endow.*, 3, 2010.
- [18] S. Guha et al., "Clustering Data Streams: Theory and Practice," in *IEEE TKDE*, 15(3), 2003.
- [19] A. Bifet et al., "Data Stream Mining, A Practical Approach", Center for Open Software Innovation – <http://www.waikato.ac.nz/research/units/cosi.shtml>.
- [20] J. Silva et al., "Data Stream Clustering: A Survey," in *ACM Comput. Surv.*, 46(1), 2013.
- [21] M. Hahsler et al., "Introduction to Stream: An Extensible Framework for Data Stream Clustering Research with R," <https://cran.r-project.org/web/packages/stream/>