Tailored IoT & BigData Sandboxes and Testbeds for Smart, Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem

Infinitech

D5.1 – Library of Parallelized Incremental Analytics - I

Lead Beneficiary	**LXS / LeanXcale**
Due Date	2020-08-31
Delivered Date	2020-10-19
Revision Number	3.0
Dissemination Level	Public (PU)
Type	Report (R)
Document Status	Release
Review Status	Internally Reviewed and Quality Assurance Reviewed
Document Acceptance	WP Leader Accepted and/or Coordinator Accepted
EC Project Officer	Pierre-Paul Sondag

HORIZON 2020 - ICT-11-2018

This project has received funding from the European Union's horizon 2020 research and innovation programme under grant agreement no 856632

Contributing Partners

Partner Acronym	Role[1]	Name Surname[2]
LXS	Lead Beneficiary	Ricardo Jiménez-Peris
LXS	Contributor	Boyan Kolev
GLA	Contributor	Richard McCreadie, Craig Macdonald, Iadh Ounis
CTAG	Contributor	Andrea Becerra
BOUN	Internal Reviewer	Can Ozturan
IBM	Internal Reviewer	Fabiana Fournier
INNOV	Quality Assurance	Dimitris Drakoulis

Revision History

Version	Date	Partner(s)	Description
0.1	2020-09-01	LXS	ToC Version
0.2	2020-09-02	LXS	Input on Introductory section and Executive Summary
0.3	2020-10-08	LXS, GLA, CTAG	Input in section 2
0.4	2020-10-13	LXS	Input in conclusions
1.0	2020-10-13	LXS	Submitted for internal review
1.5	2020-10-15	IBM, BOUN	Internal review
2.0	2020-10-16	LXS	Submitted for QA
2.5	2020-10-18	INNOV	QA
3.0	2020-10-19	LXS	Official submission

[1] Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

[2] Can be left void

Executive Summary

The goal of Task T5.2 “Incremental and Parallel Data Analytics” is to deliver a set of algorithms that can be used in the finance and insurance sector and can be considered incremental and be parallelized in order to improve the overall performance. In addition, they can be used by a data analyst to extract information as close to real-time as possible. Typical algorithms for these types of scenarios can be found in the area of frequent pattern mining, time series prediction analysis, collaborative filtering, and others. Even if they have been widely adopted from applications in the aforementioned sectors, they usually rely on static data that has been persistently stored in a datastore, and as a result, even if they can be parallelized, they cannot be considered as incremental.

As the INFINITECH platform provides an innovative data management layer that claims to overcome the inherited and existed barriers for correlating data *at-rest* with streaming data, it provides a unified data framework for integrated query processing on both types of sources. The latter makes use of a streaming engine that provides additional operators that allows this correlation of data, and relies on the basic pillars of the data management layer of the platform. In the scope of T5.2, we rely on the work that has been carried out in the corresponding tasks of WP3 that implements that layer, and on the results of T5.3 that provides *online aggregations.* By exploiting the advancements of those tasks, we are in position to re-design popular algorithms used in the finance and insurance sectors, implement them in a distributed manner so that they can be easily scaled out and serve very high rates, and by using the INFINITECH’s unified framework for data processing, we can deliver the results incrementally.

This deliverable describes how we take advantage of INFINITECH’s existing tools and frameworks in order to parallelize time series algorithms for correlation discovery and forecasting, a popular family of algorithms that is being used in a variety of use cases in the finance sector regarding risk assessment for stock or retail trading. Our library can be executed in parallel and the results are being returned incrementally. As this report summarizes the work that has been carried out during the first phase of the project (M05-M11), it will be further extended with additional types of algorithms that are popular and of high importance for finance and insurance organizations. Two additional versions of this deliverable will be released at M20 and M27 that will report the additional work that will be carried out in the corresponding phases of the project.

Table of Contents

List of Figures

Abbreviations

DDoS	Distributed Denial of Service
DL	Deep Learning
FFT	Fast Fourier Transformation
HTAP	Hybrid Transactional and Analytical Processing
iSAX	Indexable Symbolic Aggregate Approximation
ML	Machine Learning
PAA	Piecewise Aggregate Approximation
RNN	Recurrent Neural Network
SVD	Singular Value Decomposition
WP	Work Package

1. Introduction

Modern enterprises tend to use data coming from a variety of heterogeneous sources that are collected via numerous means and usually stored in a data warehouse or a data lake. Data analysts make use of sophisticated Artificial Intelligence (AI) algorithms for Machine Learning/Deep Learning (ML/DL) in order to extract valuable information that is crucial for the business intelligence of the organization. Focusing on the finance and insurance institutions, typical use cases include: online risk assessment through the correlation discovery of stocks or other finance products, online fraud detection of a finance transaction, optimal resource management of the overall portfolio of a customer, which can be either a person or another business institution, and potential identification of opportunities for investment.

Typical uses of such cases usually rely on historical data that have been imported into a data warehouse from an operational datastore or from a stream of events or other IoT data. The reason for migrating data from one data source to another, stems from the inherit barrier of performing analytics over an operational datastore, due to the competiveness of these two different types of workloads, as explained in the corresponding deliverables of task T3.1 ("Framework for Seamless Data Management and HTAP"). Moreover, performing analytics over a stream of data introduces another obstacle, as the sophisticated AI algorithms usually require the full scan of a huge amount of data that cannot be maintained in memory. As a result, modern architectures require the migration of data coming from various sources to a data lake, which will allow the data analysts to execute their algorithms in the complete dataset in a parallelized manner.

However, those algorithms, even if they can be distributed in order to exploit the level of parallelism that can increase their responsiveness, they rely on static data that has been stored in a persistent storage medium. As a result, they cannot be considered as incremental in the sense that they can be submitted once and let the data analysts get the results in a continuous manner. Even in the case of incremental algorithms, the aforementioned barriers remain.

The INFINITECH platform aims to overcome those barriers by implementing under the scope of T3.1, a data management layer that is able to first serve Hybrid Transactional and Analytical Processing (HTAP), being distributed while ensuring data consistency and transactional semantics on the same time, and allows for data ingestion in very high rates. The latter is crucial, as it allows for a stream data, often called data *in-flight,* to be directly stored in the persistent medium of the database, without having to push it first in an interim data queue, and then make use of micro-batches for data loading, as explained in the deliverables that summarizes the work that has been carried out in the scope of T3.2. As a result, there is no need to use streaming processing to enable incremental analytics, while the algorithms can be parallelized making use of the overall dataset that is available to the organization.

Task T5.2 "Incremental and Parallel Data Analytics" aims to leverage a set of typical algorithms and libraries used for artificial intelligence in the insurance and finance sectors that can benefit from the unique characteristics of the data management layer of INFINITECH, in order to re-design them and deliver parallelized and incremental analytics. A first family of such analytics has been studied and reported in this document, and will be further extended with the complete set of library of such algorithms in the next iterations of the document.

1.1. Objective of the Deliverable

The objective of this deliverable is to report the work that has been done in the context of the task T5.2, at this phase of the project (M12). This task lasts until M27, and therefore, two more versions of this deliverables will be released, extending and modifying when necessary the content of this document. As new technological advancements coming from the other technical tasks will be delivered, the work on this task has been planned to be further developed, as it relies on the basic pillars of the data management layer. The work that has been delivered during the first phase of the project (M05-M11), was mainly focused on the parallelization of a time series algorithm for risk assessment prediction, which can be also used in other domains. We studied how to deliver this algorithm in an incremental manner, by experimenting with the results of other tasks of the project. Finally, this deliverable reports on how we can deploy such types of algorithms to make use of streaming processing in order to deliver results in an incremental manner.

1.2. Insights from other Tasks and Deliverables

The work that has been carried out in the scope of T5.2 relies on the outcomes of the T2.1 (“User Stories and Analysis of Stakeholders' Requirements”) that define the overall user stories and requirements of the use cases of INFINITECH, and how the implementation can be integrated with the achievements of the other technical tasks, as defined in the INFINITECH RA. All those are part of WP2, and more precisely of T2.3 ("Specification of Enhancements to BigData & IoT Platform") and T2.5 ("Open Banking APIs, Testbeds and Data Assets Specifications"). Apart from this, WP3 gives significant input to this task, as it implements the basic technological pillars for T5.2 to rely on. WP3 provides the overall data management layer of the project. More precisely, T3.3 (“Integrated Querying of Streaming Data and Data at Rest”) implements the unified data query processing framework, which allows the correlation of streaming with batch processing. This is fundamental for the provision of incremental analytics. Moreover, T5.3 (“Declarative Real-Time Data Analytics”) provides the implementation of online aggregations that can be used by the algorithms in this task. The online aggregations allow the pre-calculation of an aggregated value, thus removing the necessity to perform a full scan on a data table to calculate that value when this is needed.

2. Parallel and Incremental algorithm for time series analytics

Nowadays, we are witnessing the production of large volumes of complex data, often in the form of time series. These may originate from various sources, e.g. financial activities, such as stock trading or bank transactions, as well as monitoring network activity or collecting data from sensors. Time series analytics allow extracting useful insights from streams of numeric data through various statistical, algorithmic, or machine learning methods.

2.1 Methods

With the goal of leveraging the inherent features of the INFINITECH data management layer, which is currently implemented under the corresponding tasks of WP3, focusing on aspects such as its increased scalability, the allowance for data ingestions at very high rates, and the online aggregations, under the scope of this task, we are currently focusing on two aspects of time series analytics: correlation discovery and forecasting methods, discussed briefly below.

2.1.1 Time series correlation discovery

Time series correlation discovery aims at finding similarities across time series, as shown in Figure 1, based on a distance metric, most commonly the Euclidean distance. This can be achieved using various methods for dimensionality reduction, e.g. singular value decomposition (SVD) [1], Fast Fourier Transform (FFT) [2], [3], wavelets [4], piecewise aggregate approximation (PAA) [5], random projections [6], as well as indexing, e.g. the iSAX (indexable Symbolic Aggregate ApproXimation) tree index [7], locality sensitive hashing through sketches [8] and more.

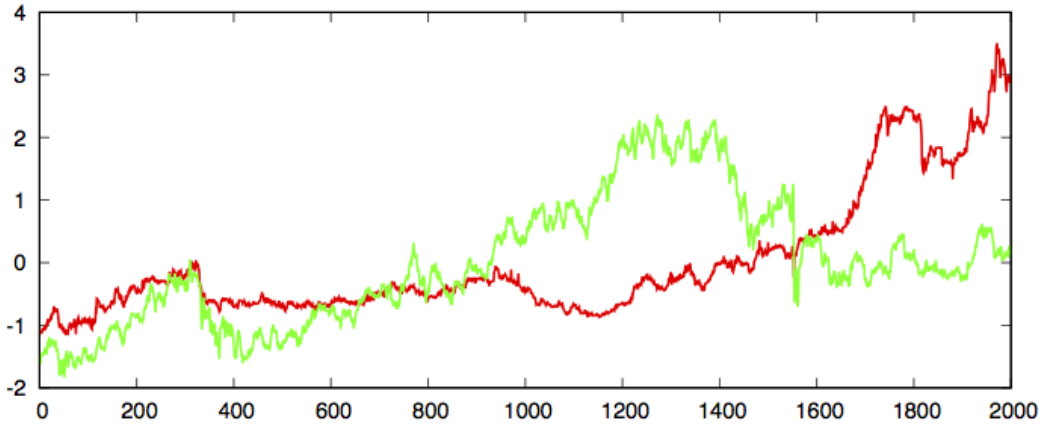

Figure 1: Example of a pair of time series that the method found to be highly correlated over the first several sliding windows of 500 time points, but not thereafter.

For the INFINITECH project, we concentrate on ParCorr, a recently introduced parallel incremental approach for fast correlation discovery over windows of time series data [8]. The method scales to millions of parallel time series, and achieves 95% recall and 100% precision. To address dimensionality reduction that nearly preserves the Euclidean distances across the latest window of time, it uses a random projection approach to compute in an incremental manner and in parallel the sketch of each time series. The sketch is a tiny structure that summarizes the time series by preserving its locality, so that simply comparing sketches provides a filtering to significantly reduce the search space for the much more expensive comparison of time series. We are focusing on that approach, taking into account the overall data management layer of the INFINITECH platform, which will exploit the use of the central data repository in order to boost the performance of this locality sensitive hashing approach. This will allow for almost real-time discovery of the top-k correlates of a given time series (see Figure 2), which further helps building a model for instantly predicting future values of the time series as a function of the values of its correlates.

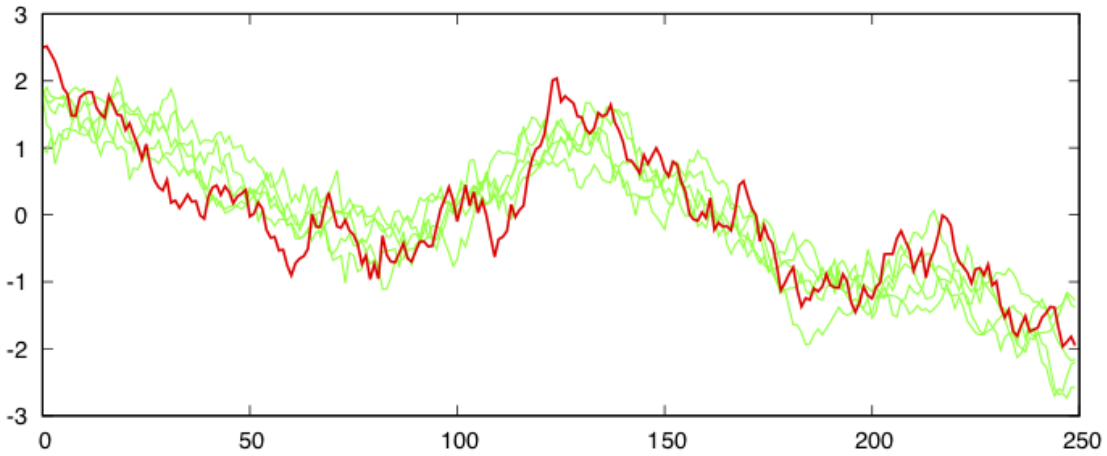

Figure 2: Example of a time series (red) and its top correlates (green) discovered by the method.

2.1.2 Time series forecasting

Statistical methods for time series forecasting [9] usually analyse the values of a single time series, often with focus on the most recent ones, to predict the next value. Simple methods, such as moving averages (MA), i.e. the average of the last few slices of time, in many cases provide good approximation for the expected value of the time series in the future moments of time. In other models of approximations that need to give more weight to the most recent values, exponential smoothing [10] would improve the forecasting accuracy. In more sophisticated scenarios, ARMA methods or deep recurrent neural networks (RNNs) can be used to capture more complex dependencies in data. Very often, all these methods work better when applied to the differences between any two consecutive points in a time series instead of directly to the time series itself, a transformation known as “differencing”.

For the INFINITECH project, we propose an efficient framework, benefiting from the features of its integral data management layer, to incrementally perform in real-time *moving averages*, *exponential smoothing*, and *differencing* of each time series towards predicting its values for the upcoming time points. Moreover, different methods can be combined to capture specific behaviour,

e.g. similarity search can be done by computing sketches on top of exponentially smoothed values and/or differencing and/or moving averages.

2.2 Potential use cases in various sectors

Time series analytics has applications in many different domains, some of which are mentioned below:

- Finance / stock trading: a pair of time series (say Google and Apple prices) that were similar before, but have ever since diverged, may represent a trading opportunity.
- Seismology: correlated signals from several different but not much distant seismic sensors may suggest that they are all related to the same seismic event.
- Network monitoring: similar traffic patterns from different sources may indicate an attempt for distributed denial of service (DDoS) attack.
- Retail / trading: future demand of a product can be approximated with statistical methods for forecasting (using the recent history of sales and/or seasonal factors), as well as through the use of discovered correlations with other indicators (such as prices and sales of other products, weather conditions, social trends, etc.).

3. Conclusions and next steps

This report documented the work that has been carried out in the scope of task T5.2 "Incremental and Parallel Data Analytics" at this phase of the project. The main objective of this task is to deliver a set of algorithms that are currently being used frequently in the insurance and finance section, relying on well-known families of machine learning implementations for correlation discovery and forecasting based on time-series prediction, clustering and collaborative filtering. The goal is to make use of these algorithms, parallelize them and make them incremental. This is enabled by exploiting the unique characteristics of the INFINITECH data management layer, and more precisely the HTAP capabilities that offers, its support for very high rate data ingestion ensuring data consistency and transactional semantics, its streaming engine and the correlation of the data *in flight* with data *at-rest,* along with online aggregates that the data management layer offers. As it can be seen, the work on these aspects is a prerequisite for the parallelization of such algorithms and their ability to become incremental, and as a fact, the main focus during this phase of the project was given on the corresponding tasks, whose work has been already reported in D3.1[3], D3.6[4] and will be delivered in the forthcoming D5.4.

This deliverable gives an overview of how we plan to parallelize algorithms for collaborative discovery and time-series prediction. Taking into consideration that gathering the relevant information regarding the specific algorithms that the pilot cases will devise is under progress at this phase, we focused on the two that we consider important for a variety of use cases. During the next phase (M12-M20) we plan to start their implementation, when all fundamental pillars required for this task are given. Therefore, more detailed information will be reported in the next iteration of this deliverable.

[3] https://drive.google.com/file/d/1Ny4c3EHhgdHOElwlCpQHJtMZjy7RR7NX/view?usp=sharing

[4] https://drive.google.com/file/d/1Rw-4eUXkF8YItJPWYlZpExgQtJBrphiF/view?usp=sharing

4. References

[1] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in timeseries databases. In Proceedings of the International Conference on Management of Data (SIGMOD), pages 419-429, 1994.

[2] R. Agrawal, C. Faloutsos, and A. N. Swami. E cient similarity search in sequence databases. In Proceedings of the International Conference on Foundations of Data Organization and Algorithms (FODO), pages 69-84. Springer-Verlag, 1993.

[3] A. Mueen, Y. Zhu, M. Yeh, K. Kamgar, K. Viswanathan, C. Gupta, and E. Keogh. The fastest similarity search algorithm for time series subsequences under euclidean distance, August 2017. http://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html.

[4] K. Chan and A. W. Fu. E cient time series matching by wavelets. In Proceedings of the International Conference on Data Engineering (ICDE), pages 126-133. IEEE Computer Society, 1999.

[5] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems (KAIS), 3(3):263-286, 2001.

[6] R. Cole, D. Shasha, and X. Zhao. Fast window correlations over uncooperative time series. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 743-749. ACM, 2005.

[7] A. Camerra, J. Shieh, T. Palpanas, T. Rakthanmanon, and E. J. Keogh. Beyond one billion time series: indexing and mining very large time series collections with iSAX2+. Knowledge and Information Systems (KAIS), 39(1):123-151, 2014.

[8] Yagoubi, D.-E., Akbarinia, R., Kolev, B., Levchenko, O., Masseglia, F., Valduriez, P., Shasha, D., 2018. ParCorr: Efficient Parallel Methods to Identify Similar Time Series Pairs across Sliding Windows. Data Mining and Knowledge Discovery, vol. 32(5), pp 1481-1507. Springer.

[9] R. J. Hyndman and G. Athanasopoulos. Forecasting: Principles and Practice. https://otexts.com/fpp2

[10] Brown, R. G. (1959). Statistical forecasting for inventory control. McGraw/Hill.