



RES. & DEV. ENGINEER – 18 MONTHS PLATFORM FOR RESEARCH AND INTEGRATION OF INCREMENTALLY OBSERVED TIME SERIES ANALYTICS

Title: Research and development engineer, platform for research and integration of incrementally observed time series analytics

Keywords: Massive time series, streaming platform operations, pattern discovery and matching in time series, time series analytics

Publication Date: 15 April 2019

Beginning Date: Q3 2019

Duration: 18 months

Contact: cedric.gouy-pailler@cea.fr

Mission

In the context of the **StreamOps** project¹, funded by the DataIA institute², CEA List is looking for a talented research and development engineer, who will be in charge of various missions around a platform for research and integration of incrementally observed time series analytics. Those missions reflect the interests of three main stakeholders: first, kernel contributors, who will be responsible for maintaining a modular and open core platform; second the algorithms contributors, who will be involved in adding state-of-the-art or cutting-edge algorithms as platform modules; third the end-user, aka scenarii developers, who will use the core platform and the algorithms to perform specific tasks on incrementally observed time series.

The missions will include:

- Contributions to the core of the platform;
- Operational concerns about the usability of the platform, such that users will be able to quickly set up the platform and run experiments from available algorithms. This will rely on combinations of containers and microservices approaches.

¹Presentation of the streamOps project in english or in french

²More information on the DataIA institute here

- Contribute some modules implementing cutting-edge algorithms, such as the ones developed at UVSQ³ or CEA⁴.
- Assist primary hand-users to set up experiments. In this regard, frontend contributions for end-users is expected (prototype).
- Data and task preparation for a benchmark

Required profile

- M2, engineering diploma or PhD in computer science.
- You have a strong background in applied mathematics/computer science (probability, statistics).
- You have very good programming skills (java, C++) and architecture design.
- You are familiar with some of these technologies: kafka, redis, influxdb.

Context

CEA is the French Atomic and Alternative Energy Commission, a leading European RTO with almost 20 000 employees and 9 centres in France. CEA is active in four main areas: low-carbon energies, defence and security, information technologies and health technologies. In each of these fields, CEA maintains a cross-disciplinary culture of engineers and researchers, building on the synergies between fundamental and technological research with around 1200 PhD students and 160 post-doctoral students. Key figures include an annual operating budget of more than 5 billion euros, 211 high-tech spin-offs since its creation, 762 priority patents filed (that reinforce a portfolio of nearly 6,300 active patent families at the end of 2017) and 422 on-going EU projects in 2017. CEA was recently ranked as one of the world's most innovative research institution (<https://clarivate.com/top100innovators>). 80% of the CAC40 (benchmark French stock market index) companies have set up bilateral agreements with CEA for the provision of targeted innovation and technology transfer services that represents more than 600 industry partners to date. In this project, CEA will participate through its Technology Division, also known as CEA Tech.

CEA Tech's three institutes (Leti, Liten, and List) develop a broad portfolio of technologies for ICTs, energy, and healthcare. CEA Tech leverages a unique innovation-driven culture and unrivalled expertise to develop and disseminate new technologies for industry, effectively bridging the gap between the worlds of research and business. In recognition of this unique positioning, CEA Tech has been awarded the French "Carnot" label, which translates into an investment of 25% of their 600 million euros yearly budget into scientific resourcing actions, to generate major technology breakthroughs. Providing a unique combination of expertise and research facilities, that encompass all required scientific and technology dimensions (software, embedded and sensor systems developments towards applications aiming at defense, security, energy, nuclear, environmental, and health), CEA LIST is part of the dynamic and challenging environment of the University Paris Saclay – the largest French scientific

³A Scalable Engine for efficient and expressive Time Series Classification [JZT]

⁴Streaming binary sketching based on subspace tracking and diagonal uniformization [MSG17]

cluster with 60 000 students – and features high-standard ICT infrastructures, as well as an extensive portfolio of software assets (CAD tools, simulation & data mining tools, large energy data records from previous projects). CEA LIST counts more than 700 researchers dedicated themselves to software-intensive systems innovation. The Institute signs 200 research contracts per year, filed 213 patents and has a turnover of 62 million euros.

The Data Intelligence Service (SID) is a reference unit focused on the development of smart algorithms and data processing. This applied mathematics laboratory addresses all the chain from raw data processing to data mining and decision making. The SID is composed of more than 50 engineers and researchers (most of whom with a PhD) who are working with academics as well as industrial partners to develop new and innovative algorithms. In particular, the SID develops mathematical approaches and associated software platforms for signal processing, data visualization and clustering, machine learning, integration of expertise and statistical model building. With 3 to 5 PhD defense per year and more than 130 international publications (between 2008 and 2014), as well as numerous industrial partnerships, the SID is at the junction between academic works and real world problem solving.

References

- [BHKP10] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa: Massive online analysis. *The Journal of Machine Learning Research*, 11:1601–1604, 2010.
- [FBH14] John Forrest, Matthew Bolanos, and Michael Hahsler. Introduction to stream: A Framework for Data Stream Mining Research. 2014.
- [JZT] Karine Zeitouni Jingwei Zuo and Yehia Taher. Exploring interpretable features for large time series with SE4TeC. In *Proceedings of EDBT 2019*.
- [MSG17] Anne Morvan, Antoine Souloumiac, Cédric Gouy-Pailler, and Jamal Atif. Streaming Binary Sketching based on Subspace Tracking and Diagonal Uniformization. *arXiv:1705.07661 [cs]*, May 2017.
- [Mut05] S Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers Inc, 2005.
- [TYW18] Huangshi Tian, Minchen Yu, and Wei Wang. Continuum: A Platform for Cost-Aware, Low-Latency Continual Learning. In *Proceedings of the ACM Symposium on Cloud Computing, SoCC '18*, pages 26–40, New York, NY, USA, 2018. ACM.