EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Internship Offers 2018-2019

For the latest information, please visit https://research.euranova.eu.
6th December 2018

CONTENTS

EURA NOVA

INTRODUCTION

EURA NOVA is a data-driven Belgian company founded in September 2008 and located in Brussels, Marseille, and Tunis. Our mission is simple: bring life to our customers' great ideas, by offering best-in-class services in data science, software engineering, and data architecture. To do so, we invest significantly in in-house expertise and state-of-the-art knowledge. In line with this course of action, we offer academic programs in collaboration with universities. These offers include bootcamps, master theses topics, research internships, and PhDs topics. See below for details.

OUR INTERNSHIPS OFFERS

This document presents internships topics supervised by our software engineering department or by our research & development department. Each project is an opportunity to be actively involved in the development of solutions to address tomorrow's challenges in ICTs and to implement them today.

The students will work in a dedicated **international** team of engineers **with diverse expertise in machine learning, graph theory, artificial intelligence, high performance computing, etc.**

They will keep EURA NOVA informed of the project advancement and share their ideas and challenges using the in-house knowledge management tool. We value continuous learning and teamwork. We love to have a good time together. For more information on our R&D activities, please visit our website at https://research.euranova.eu.

HOW TO APPLY

When you have gone through our internships offers, pick your favourite three. Draft a short text for each one, stating why you find it interesting and what you would do about it. Send us this statement, along with your CV at career@euranova.eu.

RESEARCH

APPRENTICE DATA SCIENTIST

Context

According to the Harvard Business Review magazine, data scientist is the sexiest job of the 21st century. Here, at EURA NOVA, we also think it is the coolest job. But what are we talking about? We are talking about ingesting diverse data sources, running distributed machine learning algorithms, visualising big graphs and so on. Sounds great, right?

Business Opportunity

At EURA NOVA, we leverage data to extract valuable insights for our clients. There are so many techniques to test and there is so much knowledge to digest. We hope that by sharing our knowledge with you, you will help us explore datasets much deeper.

Contribution

During the internship, you will be part of the data science team and will work in a stimulating environment. You will participate in data science projects using cutting-edge tools and techniques from the big data and machine learning domains.

BUILDING GRAPH EMBEDDINGS TO PREDICT SPARK PERFORMANCE

Context

Dealing with big data has become "business as usual" for companies and, to deal with it, distributed frameworks such as Spark have become widely used. But a framework like Spark has a lot of possible configurations, leading to different performance results. Hence the RD centre of EURA NOVA would like to predict the performance of any Apache Spark application. A Spark application is constituted of a series of tasks. The Spark engine builds a Directed Acyclic Graph (DAG) to represent the dependencies between the tasks. At EURA NOVA, we enrich this DAG with some other information to obtain a Directed Task Graph (DTG).

In order to train a performance model to predict the performance of a Spark application, one needs to define features of such an application. A possibility could be to represent the graph as a vector using a graph embedding technique. The goal of the embedding process consists in having the vectors of similar applications close to each other in the projected space.

Technologies

Current implementations of some common graph embedding techniques exist in Python, but other programming languages or frameworks can be used as well.

Objectives

This internship will focus on developing a graph embedding technique for our DTG, to be able to predict the performance of a Spark application. Moreover, a performance measure of the embedding technique should be developed as well. Two vectors should be close to each other if the two corresponding applications are similar. The goals of this internship are:

- To get familiar with current graph embedding techniques
- To get familiar with the performance prediction of a Spark application
- To develop a graph embedding approach suitable in our setting
- To evaluate the performance of the chosen embedding approach

ANALYSING KEY FEATURES OF A SPARK WORKLOAD

Context

Dealing with big data has become normal for companies and distributed frameworks such as Spark have become widely used. But a framework like this has a lot of possible configurations, leading to different performance results. Hence we would like to predict the performance measure of any Spark application based on the features of the input data sent to the application.

Technologies

You will work with Apache Spark with the Dataframe API. You will need Scala to fully understand how Spark works internally. You might choose to use some JMX endpoints that Spark provides to record metrics, in which case Java would be needed.

Objectives

This internship will focus on one of the following groups of applications in Spark: streaming, machine learning, SQL, and graph processing. The goals of this internship are:

- to define the features to use to predict the performance measure of any Spark application
- to record the performance measure of Spark applications
- to build a performance predictor and assess its accuracy

STREAMING PLATFORMS COMPARISON: A FAUSTIAN TALE

Context

There is a growing need for real-time applications, i.e. some data processing in which we can respond to an action immediately after it happened. This is the case, for example, in fraud detection banking systems, or in real-time advertising campaigns based on a stream of user clicks. To develop a streaming application, we rely most of the time on tools such as Kafka, Spark, or Flink. To set up such tools, some amount of data engineering is needed to get started, and this stage takes some time. In the exploratory phase of a project, being able to reduce this setup time would allow us to focus on the application itself.

Technologies

Faust (https://github.com/robinhood/faust) is a stream processing library developed in Python 3.6, importing the ideas of Kafka Streams to Python.

Objectives

During this internship, you will:

- explore the Faust library and its possible use cases
- compare this library to Kafka Streams in terms of capabilities, ease of use and performance

A BENCHMARKING TOOL FOR DISTRIBUTED PROCESSING ENGINES

Context

Dealing with big data has become normal for companies and distributed frameworks such as Apache Spark or Apache Flink have become widely used. Sometimes data flows continuously and data needs to be handled on the fly. This is the case, for example, in fraud detection banking systems, or in real-time advertising campaigns based on a stream of user clicks. Spark streaming and Flink can be part of the solution for such cases, but frameworks like these have a lot of possible configurations. Hence, reporting metrics about these frameworks can help better understand what configuration parameters to change. In cases like this, Apache Kafka is often chosen as the input and the output of the distributed processing framework.

In this internship, we will focus on the recording of the latency of each event handled by the system. You will build a benchmarking tool to plug into any existing distributed processing framework to allows for quick and easy deployment at the client's premises. Then EURA NOVA will be able to collect the latencies of the events coming from any client's workload.

Technologies

You will work with either Spark streaming or Flink, and Kafka.

Objectives

The objectives of this internship are:

- to develop a benchmarking tool to record event latencies of the messages treated during any workload of a distributed processing framework,
- to run the benchmarking tool for different use cases and report the results.

BENCHMARKING RESOURCE CONSUMPTION OF DISTRIBUTED PROCESSING ENGINES

Context

Dealing with big data has become normal for companies and distributed frameworks such as Apache Spark or Apache Flink have become widely used. Sometimes data flows continuously and data needs to be handled on the fly. This is the case, for example, in fraud detection banking systems, or in real-time advertising campaigns based on a stream of user clicks. Spark streaming and Flink can be part of the solution for such cases, but frameworks like these have a lot of possible configurations. Hence, reporting metrics about these frameworks can help better understand what configuration parameters to change. In cases like this, Apache Kafka is often chosen as the input and the output of the distributed processing framework.

In this internship, we focus on the resource usage of the framework: how much CPU and how much memory are used by the job. You will build a benchmarking tool to plug into any existing distributed processing framework allows for quick and easy deployment at the client's premises. Then EURA NOVA will be able to collect the resource consumption of any client's workload.

Technologies

You will work with either Spark streaming or Flink, and Kafka.

Objectives

The objectives of this internship are:

- to develop a benchmarking tool to record CPU and memory consumption of any workload run in a distributed processing framework,
- to run the benchmarking tool for different use cases and report the results.

APPRENTICE DATA SCIENTIST FOR DIGAZU

Context

digazu[1] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Objectives

This internship will focus on defining and building end-to-end use cases using digazu capabilities:

- Collect data in real time
- Apply transformations on streamed data
- Develop real-time dashboards
- Apply data science models in real time

This will allow you to get to know cutting-edge big data technologies, as well as to get familiar with data architecture concepts and challenges. It will also introduce you with the end-to-end process of leveraging big data use cases.

[1] https://digazu.eu/

APPRENTICE DATA ARCHITECT FOR DIGAZU

Context

digazu[2] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Objectives

This internship will focus on analyzing the market of data integration platforms and on comparing the capabilities of digazu with its main competitors, such as Informatica, Talend, Attunity…

This will allow you to get to know cutting-edge big data technologies, as well as to get familiar with data architecture concepts and challenges. It will also help you gain knowledge on the big data landscape and trends.

[2]https://digazu.eu/

DATASETS MANAGEMENT FOR RESEARCHERS IN DIGAZU

Context

digazu[3] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Objectives

When conducting research projects, researchers often need to find, extract and integrate open datasets. This internship will focus on gathering such datasets and make usage of digazu to ingest and integrate the data so that it can be used easily by researchers. This internship will also cover the industrialization of data science models developed by researchers, within digazu.

This will allow you to get to know cutting-edge big data technologies, as well as to get familiar with data architecture concepts and challenges. It will also introduce you to the end-to-end process of leveraging big data use cases.

[3] https://digazu.eu/

ENGINEERING

INFRASTRUCTURE AUTOMATION / CONFIGURATION MANAGEMENT

Context

The important growth of distributed systems and the emergence of the Dev-Ops philosophy has blurred the separation between operations (i.e. 'running the server') and development (i.e. 'building the app') in many companies.

This has led to the emergence of new tools to handle infrastructure automation and predictable deployment. Tools span from configuration management / deployement automation (Puppet, Chef, Ansible, Salt, etc) to resource management (Mesos, Yarn), functional package management (Nix) and app container management (Docker, Lmctfy).

EURA NOVA wishes to further explore how these new tools can facilitate deployment and maintenance of both its internal and external projects.

EURA NOVA is deploying its current generation of tools using Ansible to maintain the machines and Docker to build and deploy the services on top of it. All the steps are automated and reproducible. We would like to push further in the construction of our infrastructure.

Objectives

This internship has several objectives:

- participate in the deployment of the current generation of EURA NOVA's internal services on its new production infrastructure, which will give you the opportunity to familiarize yourself with some of the aforementioned technologies (Docker, Ansible, Gitlab CI).
- participate in the deployement of an artifact registry, i.e. a service where developers will be able to deploy releases of jar, npm packages, Docker images, etc.
- design and implement the architecture of a fully-automated build and deploy system based on Gitlab CI, Docker, and the artifacts registry.

FRONTEND DEVELOPMENT

Context

Both for its internal tools and in the context of its R&D projects, EURA NOVA often needs to build pretty and efficient web applications quickly in coordination with backend developers, data scientists, etc.

We are looking for a front-end developer to join the team and work on these projects.

Proficiency

Though EURA NOVA strives to work with agility, it also wishes to build durable products: its developers try to write clean code. If you are proficient with a technology, your experience will be counted on to improve our practices. If you are not, you must be willing to strive to do better everyday.

Technologies

You would work with the following technologies:

- Git and Gitlab for collaboration,
- HTML, CSS and frameworks such as Bootstrap or Material-UI
- Javascript with ReactJS as the main brick for building your UIs,
- visualization libraries such as D3.js or vis.js
- technologies to interact with APIs: REST, websockets, Thrift, etc.

Objectives

The objective of the internship is to join the pool of front-end devs and take an active part in our different projects where a web application is needed.

COMPARISON OF FLINK EXECUTION MODES

Context

Flink is a distributed computing framework, so one of its main feature is to execute a task of a job on a specific machine of a cluster. How the machine is selected is up to an orchestrator. In the cluster orchestrator world, Kubernetes has become the standard. Flink jobs can be executed in two modes either relying on Kubernetes to assign the task to a machine or relying on the internal Flink orchestrator process.

Technologies

Apache Flink, Kubernetes, Apache YARN

Objectives

Benchmark execution time and resource consumption for the following execution modes:

- Internal Flink orchestrator on YARN (Default)
- Internal Flink orchestrator on Kubernetes
- Kubernetes orchestrator

Using the comparison of the results, the goal is to challenge the assumption that one mode is better than the others, and find out which one.

REFERENCES