Job Offer: Big Data Engineer (Apache Spark and Scala)

Summary

You can help to make the world a better place. One spark-submit at a time.

- **Institution:** École Polytechnique
- **Location:** Palaiseau, France
- **Required Skills:** Software Development | Apache Spark (Scala) | Big Data
- **Bonus Skills:** Python | Machine Learning | TDD
- **Languages:** English and/or French
- **To apply:** https://stackoverflow.com/jobs/153458/

Who are we?

The Data Science Initiative research team is seeking for a Software Engineer to work on one of our main projects. We are working on pattern detection using machine learning techniques on one of the world's largest health database. This database records every health related transaction for more than 60 millions people, resulting in hundreds of Terabytes of data. The project outcome will have a broad societal impact, either from medical or economical viewpoints.

Our team is composed of several enthusiast mathematicians (international researchers and PhD students) and data-driven software engineers. It is a team with a big ambition, and a lot of room for creative solutions.

What we do

Right now, we are using a simple stack based on an HDFS - Spark cluster. We are also developing an open source C++/Python library called tick (you can check its Github page) that features several homemade machine learning algorithms for longitudinal data analysis.

The main workflow is quite standard:

- Cleaning: from a production SQL database, we format the data in a (very) large denormalized table;
- Featuring: we compute a lot (a lot) of features based on this table to create a large matrix;
- Learning: we feed our homemade machine learning algorithm with this matrix and estimate a model.

Typical day

- You develop a lot of stuff in Spark and try different design to improve our performances using any kind of kick-ass technology, all of that using a slick git workflow;
- You walk across the campus and cross some students, teachers and horses;
- You can play with a cluster to get better performance (or just to learn). You can also use large machines to run your code;
- You can use the sport facilities on campus during if you want;

- You are also part of every decision regarding the evolution of our architecture (such as numerous NoSQL debates for instance);
- You are also looking for new libraries, releases, papers, conferences, literally anything that's linked to your field of interest. The campus library is fully stocked, and you can order what's missing.

Who are you?

- Curious is your middle name.
- GitHub and StackOverflow are your favorite social networks.
- You know why you're using Eclipse instead of IntelliJ or the other way around.
- You spent several years studying computer science or mathematics and more time playing with algorithms. You acquired good programming skills and know about clean code, tests and versioning.
- You've seen things. Weird implementations and hairy publications. Everywhere.
- You dream in Scala or Python (and you don't talk about nightmares).
- You've heard of functional programming, but you're not supposed to be killer at it.
- Also, you know about the machine learning ecosystem and everything revolving around distributed computing.
- You can survive at least a month without water or food on Unix and you've picked your stand between emacs and vi.

Bonus

You are either:

- A killer in Scala;
- A hardcore Spark developer;
- Fluent in python;
- An experienced sysadmin.

You will be able to:

- Work in a highly challenging intellectual environment;
- Wear sandals with socks (but why would you do that?);
- Benefit from the campus infrastructure, which means that you could practice horse riding, golfing and water-polo, unfortunately not at the same time;
- Use a lot of different technologies and try any that looks appealing to you;
- Have access to adequate conferences, books and trainings if needed.