## JOB SPECIFICATION
## FOR APPLICATION

**Publication date:** March 2019

**Job title:** Post doctoral in Mutable Shared Data for Serverless Computing Infrastructures

**Location :** Evry

**School/department :** TELECOM SudParis –Département INFormatique (INF)

**Post of hierarchical superior :** Maître de conférences en informatique

**Categories or trades of agents who can apply :** II – P or PhD student

**Job Category :** II - P

**12-month contract**

## MISSIONS AND CONTEXT:

**Background**
The recently funded H2020 CloudButton project [1] aims to democratize big data by overly simplifying its programming model with the help of serverless technologies.The core idea is to tap into stateless functions to enable radically-simpler, more user-friendly data processing systems. Average users of the cloud do not want to spend hours understanding complex analytics stacks (e.g., Spark |2], Yarn [3], or Ignite [4]), and to struggle with the choice of instance types, cluster sizes, etc. What they want is just a simple interface to execute their optimized, single-machine code in parallel. CloudButton is the technological response to this emerging need. To demonstrate impact, the project targets two strategic settings with large data volumes and diverse analytics requirements: bioinformatics (genomics, metabolomics) and geospatial data (LiDAR, satellital).

## ACTIVITIES AND TASKS :

**Objectives**
Translating a single-machine concurrent code to a massively-parallel serverless infrastructure requires at core to support mutable shared data [5]. The main objective of this postdoc position is to specify and implement such a storage system for the CloudButton stack. Serverless computing infrastructures deliver short-lived functions where computation quickly scales up and down. To cope with this programming model, the considered layer needs to be auto-scalable and shared data ephemeral, lasting only for the duration of the serverless function calls [6]. In particular, this situation may require to tackle the challenging problem of co-locating data together with

computation and thus to operate the storage layer at large-scale during a brief amount of time. In addition, the storage layer of CloudButton should simplify the transitioning from single-machine code to serverless infrastructure. To this end, we envision that its interface consists of distributed shared objects implemented in a server tier and callable remotely by (object-oriented) serverless functions. To improve code modularity, objects are composable. For performance, the storage layer may also split objects transparently to the serverless functions and supports objects with various consistency levels and different degrees of replication.

**Work Plan**

For starter, this project will rely on the Infinispan data grid [7] developed by RedHat, a CloudButton partner, and the contributions made in the Creson framework [8]. A tentative agenda may consist in the portage of an existing machine learning Java library with the help of this framework, and its evaluation in practice using a standard data analytics workload.

**References**

- [1] http://cloudbutton.eu, retrieved Feb. 2019
- [2] Matei Zaharia, Reynold S. Xin, Patrick Wendell et al. , Apache Spark: a unified engine for big data processing, CACM, Nov. 2016
- [3] Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas et al, Apache Hadoop YARN: yet another resource negotiator, Oct. 2013
- [4] Shamim Bhuiyan, Michael Zheludkov and Timur Isachenko, High Performance in-memory computing with Apache Ignite, 2017
- [5] Raul Castro Fernandez, Panagiotis Garefalakis and Peter R. Pietzuch, Java2SDG: Stateful big data processing for the masses, ICDE, May 2016
- [6] Ana Klimovic, Yawen Wang, Patrick Stuedi et al. Pocket: Elastic Ephemeral Storage for Serverless Analytics, OSDI Oct. 2018
- [7] http://infinispan.org, retrieved Feb. 2019
- [8] Pierre Sutra, Etienne Rivière, Cristian Cotes et al., CRESON: Callable and Replicated Shared Objects over NoSQL, ICDCS, June 2017.

**TRAINING AND SKILLS :**

**Level of training and / or experience required:**
- PhD in Computer Sciences or related fields

**Essential skills, knowledge and experience:**
- Excellent academic record
- Strong background in distributed systems / database / algorithms
- Good developer and experimenter at large scale
- Speak English

**Desirable skills, knowledge and experience:**
- Knowledge of object-oriented programming (notably Java) is a plus
- To speak French

Candidate information on the processing of personal data: https://bit.ly/2QeOZhl