Project Proposal Form
Industrial Project course 234313
Computer Science Department, Technion

Serverless for Big Data processing in the cloud

Serverless computing is an emerging technology that has the potential to radically change the way big data processing is done. Most big data analytic flows can benefit from the serverless platforms, starting with simple cases of processing object storage data, to more complex data preparations for AI frameworks, like TensorFlow.

To address the challenge of how to easily integrate serverless computing, without major disruptions to your system or code rewrites, the IBM Cloud Functions team and IBM Research developed the PyWren-IBM-Cloud framework. Based on the open source PyWren project, this new framework offers a brand new "push to the cloud" experience for the users. It allows them to focus strictly on writing their Python code, while PyWren deploys the code as a serverless action to IBM Cloud Functions, monitors its execution, and runs it with a large amount of parallelism.

Overall goals

1. Demonstrate embarrassing parallel algorithms with PyWren over IBM Cloud. The first candidate is Monte Carlo simulations.
2. Can we run NumPyWren (https://github.com/Vaishaal/numpywren) over PyWren-IBM Cloud?
3. Facebook https://github.com/facebookresearch/fastText (https://en.m.wikipedia.org/wiki/Hyperparameter_optimization) with PyWren and serverless? https://fasttext.cc/docs/en/supervised-tutorial.html
4. Students may decide to demonstrate other interesting embarrassing parallel problems, for example to choose from http://www.cs.nthu.edu.tw/~ychung/slides/para_programming/slides3.pdf or http://www.cs.iusb.edu/~danav/teach/b424/ or https://www.cs.fsu.edu/~engelen/courses/HPC/Algorithms1.pdf and so on. Students have a freedom to choose other problems, but need to verify with project supervisor
5. Students will add code improvements to PyWren and contribute code directly against open source repository. We will discuss those steps during course of the project

Phase 1 - setup

IBM Cloud Academic Initiative

1. Register to IBM Cloud https://www.ibm.com/cloud/ using your academic email. Registration is free of charge, no need to provide payment method.
2. Register to IBM Academic Initiative
   a. Follow https://ibm.onthehub.com/WebStore/OfferingDetails.aspx?o=bb3528b7-2b63-e611-9420-b8ca3a5db7a1
   b. Use your student email account
   c. You will get an access code, that you need to copy-paste into IBM Cloud: navigate to profile, billing, promote code and paste the code
3. From the IBM Cloud catalog, choose:
   a. IBM Cloud Object Storage
   b. Watson Studio
   c. Cloud Functions – **follow all the steps, including CLI installation. Make sure you can run "hello world" example with Python**

PyWren Code, IDE setup

1. Checkout PyWren for IBM Cloud https://github.com/pywren/pywren-ibm-cloud (you need to have a github account, git tool, etc.)
2. Setup your preferable IDE for Python: Eclipse / IntelliJ, etc.
3. Create new Python project and import pywren-ibm-cloud
4. Follow documentation from the https://github.com/pywren/pywren-ibm-cloud
5. Make sure you can run end-to-end example

You are done!
Summary of the Phase 1

1. You created IBM Cloud account with 6-month of free academic usage
2. You installed IBM Cloud Object Storage Service
3. You created IBM Cloud Function service and managed to run hello world action with Python
4. You installed local IDE and imported PyWren project from github
5. You managed to run local PyWren against IBM Cloud Functions

Phase 2 – What is next

Any embarrassing parallel algorithm can benefit from PyWren execution model.

More to read

1. Paper provided – "Serverless Data Analytics in the IBM Cloud". **Please don't distribute the paper, for your personal use only**
2. http://shivaram.org/publications/pywren-socc17.pdf
3. Apache OpenWhisk https://openwhisk.apache.org

4.