Astronomical Data Analysis Software and Systems XXVIII
ASP Conference Series, Vol. 523
P.J. Teuben, M.W. Pound, B.A. Thomas, and E.M. Warner, eds.

Working with the Hubble Space Telescope Public Data on Amazon Web Services

Ivelina G. Momcheva

Space Telescope Science Institute, Baltimore, MD, USA; `imomcheva@stsci.edu`

Abstract. In May 2018 STScI announced that ~110 TB of archival observations from the *Hubble Space Telescope* are available in cloud storage on Amazon Web Services. This tutorial provides an introduction to accessing this dataset and to AWS cloud computing in general for users who have not previously used cloud resources. We demonstrate how to access the Hubble Public Dataset on AWS in order to carry out a variety of tasks. Participants are introduced to the `astroquery.mast` and `boto3` Python client libraries. We demonstrate operations with the data such as retrieving, displaying and running analysis on single images. We then show how to scale up analysis through serverless computing. Finally we browse some advanced capabilities such as logging, price estimation and machine learning. The tutorial is geared toward novice AWS users. Intermediate Python knowledge is strongly recommended.

1. Introduction

The *Hubble Space Telescope* has undeniably expanded our understanding of the universe during its 28 years in space so far, but this is not just due to its superior view from space. One of the major advantages to *Hubble* is that every single image it takes becomes public within six months (and in many cases immediately) after it is beamed back to Earth. The treasure trove that is the Hubble archive has produced just as many discoveries by scientists using the data "second hand" as it has from the original teams who requested the observations.

In May, 2018 we announced that 110 TB of *Hubble's* archival observations are available in cloud storage on Amazon Web Services (AWS) which provides unlimited access to the data right next to a wide variety of computing resources. These data consist of all raw and processed observations from the currently active instruments: the Advanced Camera for Surveys (ACS), the Wide Field Camera 3 (WFC3), the Cosmic Origins Spectrograph (COS), the Space Telescope Imaging Spectrograph (STIS) and the Fine Guidance Sensors (FGS). The data on AWS[1] are kept up to date with the data held in the Mikulski Archive for Space Telescopes (MAST). New and reprocessed data are updated on AWS within 20 minutes of them being updated at MAST. The combination of cloud computing with one of the highest value dataset in astronomy has the potential to yield new scientific discoveries by allowing users to do large scale data analysis and utilize cloud services.

[1] `https://registry.opendata.aws/hst/`

We have created a tutorial which demonstrates how to access the Hubble Public Dataset on AWS in order to carry out a variety of tasks. Access to the data is provided through a custom extension to the `astroquery` Python library – `astroquery.mast` – and the AWS client Python library `boto3`. The tutorial introduces participants to the basic functionality. It further demonstrates operations with the data such as retrieving, displaying and running analysis on single images. It then shows how to scale up the analysis to hundreds on images through AWS Lambda serverless computing.

2. Learning Objectives

The primary objective for the tutorial is to get users started with the *HST* AWS dataset. This goal will be accomplished with the following activities:

- Learn about the Hubble Public Dataset on AWS
- Learn about `astroquery.mast` and `boto3`
- Download the data
- Visualize the data and carry out data analysis
- Create a function that can be parallelized
- Run a function using Lambda (serverless computing)
- Basic exploration of machine learning resources

Even though cloud computing is now widely used in industry applications of big data, the uptake of this technology in the astronomical community has been slow. A secondary objective of the tutorial is to introduce the participants to several key services provided by cloud computing platforms including on-demand computational resources, storage, serverless compute and machine learning capabilities. Along with this, participants learn about estimating costs, logging activity and connecting different cloud services to execute a task.

The secondary learning objectives are accomplished with the following tasks:

- Creating an AWS account and logging in
- Starting Amazon machine images (AMIs) image and connecting to it
- Creating a Docker container based on a template
- Writing a Lambda function
- Different types of cloud storage, creating a new bucket, connecting to them
- Logging activity
- Prototyping computations and estimating cost

3. Tutorial

The tutorial is based on blog posts in the Mast Labs blog.[2] A basic/intermediate knowledge of Python is a prerequisite for this tutorial. Users need a laptop and an internet connection. The contents of the tutorial are hosted in a public Google document:

https://tinyurl.com/adass2018-aws

Figure 1 shows the final output from the Lambda task. Lambda is serverless compute, where the user specifies the software but not the hardware. It is well suited for small tasks that are easily parallelize-able. In our example we carry out source detection using `sep`[3] and produce a gray-scale PNG file of the image with red circles at the positions of all sources. The example case executes in seconds on one hundred images and the cost is covered under the free Lambda tier.

Users are encouraged to work through the tutorial on their own time. Comments and suggestions for improvements are welcome at imomcheva@stsci.edu.

Even though AWS offers a free tier for most of its services, some of the tutorial requirements are not covered under it and a valid credit card is needed order to create an AWS account even if no charges are incurred. Going through the materials in the tutorial cost less than 5 cents to run and at the end of the tutorial we show users how to clean their workspace so no further charges are incurred. For larger scale computation AWS does offer cloud credits for research grants[4] through a competitive process. Starting in HST proposal cycle 26 (2018), STScI established a new type of proposal specifically utilizing this dataset which provides up to $10,000 for AWS services. Additionally, all archival and general observer programs can request AWS finds in their budgets.

4. Conclusion

The tutorial presented here introduces users to the *HST* data on AWS and to cloud computing in general. Cloud compute combined with astronomical data offers new capabilities for discovery. Some of the services offered by cloud providers (e.g., Lambda) do not have easy analogues that can be run locally and with so little effort. We envision many more use cases for cloud computing in the near future.

Acknowledgments. This tutorial was supported by AWS who provided cloud credits for all participants. We thank the tutorial helpers C. Brasseur, P.-L. Lim and N. Miles.

[2]https://mast-labs.stsci.io/

[3]https://github.com/kbarbary/sep, a Python wrapper around the core algorithms of Source Extractor

[4]https://aws.amazon.com/grants/

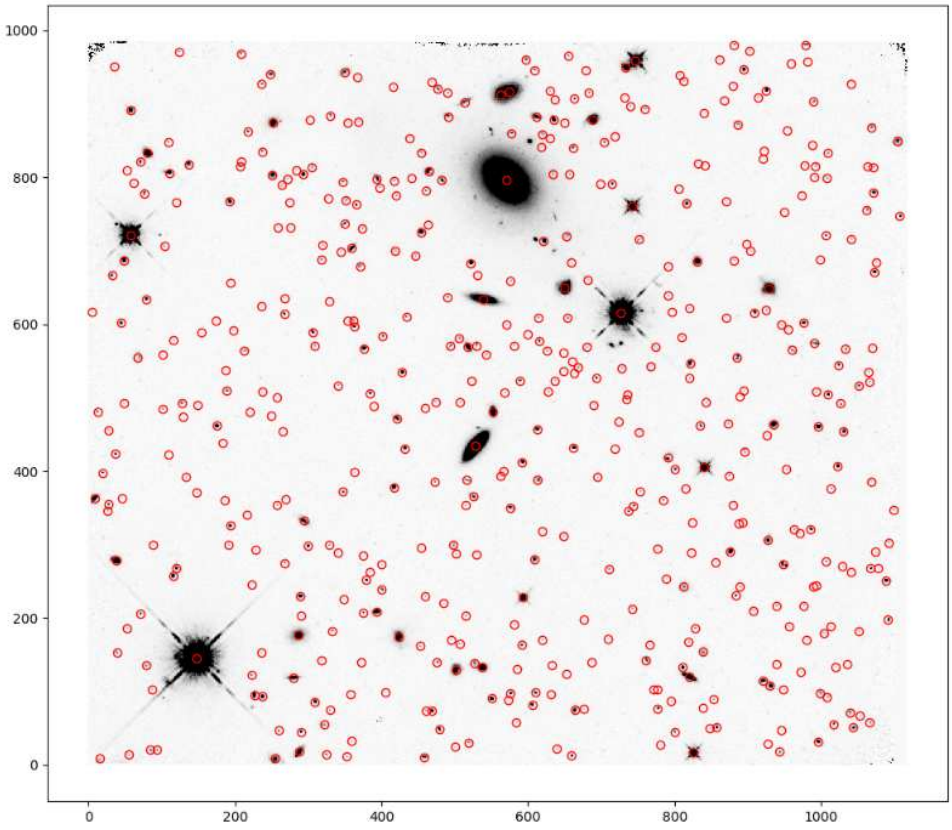

Figure 1. Sample output from the AWS Lambda function executed as part of the tutorial. The background is a grayscale *HST* WFC3/IR image. Red circles mark the position of detected sources.