How to get data?

By Jingbo Shang (**shangjingbo**)

# Preface

Downloading data for The SpaceNet Challenge will be a little different from how Topcoder usually provides you data for a match. That being said, it will be a straightforward and painless process if you follow this **“how to”** guide. In total, if you already have a linux, mac machine, or AWS CLI for Windows, the process should take you ~10 minutes to get to the data!!!

Follow our guidance in this document, which includes:

* For batch downloading, a Python script to expedite the data download.
* For single file downloading, command-line commands are provided.
* Step-by-step configurations with screenshots for better illustration. (This is why the document is a little long.)

Note that downloading this data will cost you **only** about $0.45 to $1.35. The data is $0.09 per GB to download. So although you will be asked to provide a credit card, there are only ~5GB to 15GB of data (depending on whether you download the full data set or just the minimal subset) to download for this match to compete.

Note that for “Data Transfer OUT From Amazon S3 To Internet” gives you:

* First 1 GB / month $0.000 per GB
* Up to 10 TB / month $0.090 per GB
* If you are a new AWS customer, for up to 12 months following your AWS sign-up, you are given a free usage tier of 15GB. Once your usage expires or your application use exceeds the free usage tiers, you simply get pay-as-you-go service rates (see each service page for full pricing details). Restrictions apply; see offer terms for more details.

In the future, Topcoder will run more SpaceNet Challenges with larger datasets that will follow this same download process.

Table of Contents

[Preface 1](#_Toc467074533)

[AWS account 3](#_Toc467074534)

[Set up a new user in your AWS account 4](#_Toc467074535)

[Get a Linux or Mac OS 6](#_Toc467074536)

[Install AWS CLI 6](#_Toc467074537)

[Mac and Linux 6](#_Toc467074538)

[Windows 6](#_Toc467074539)

[Configure AWS CLI 6](#_Toc467074540)

[Download Single File 6](#_Toc467074541)

[Download Batch Data 7](#_Toc467074542)

[Minimal data set 7](#_Toc467074543)

# AWS account

If you have an Amazon shopping account, you already have an AWS account implicitly.

[Amazon Web Services](https://aws.amazon.com/) (AWS) , a subsidiary of Amazon.com, offers a suite of cloud-computing services that make up an on-demand computing platform.

In the top-right corner, there is a “**Sign In to the Console**” button as the following.



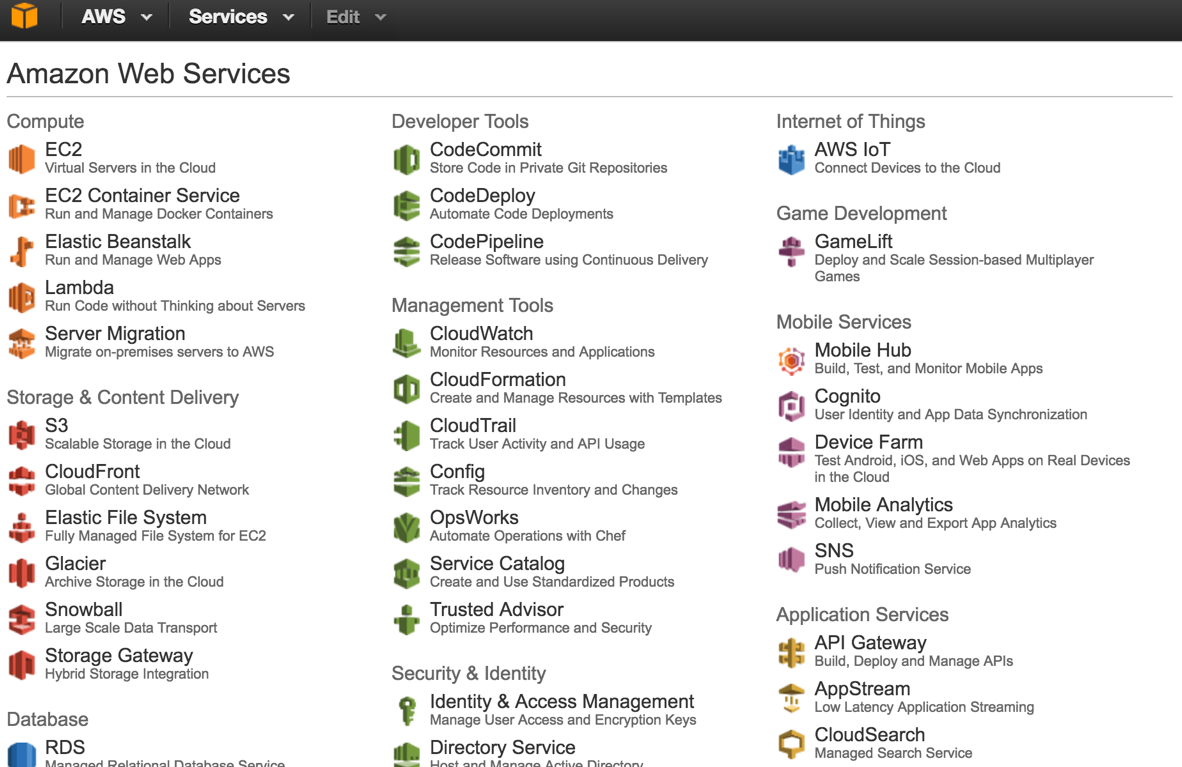
Amazon Web Services uses information from your Amazon.com account to identify you and allow access to Amazon Web Services. Therefore,

* If you already have an amazon account (e.g., for online shopping), you can simply use the same.
* Otherwise, please register a new account through [registration website](https://aws.amazon.com/resources/create-account/).

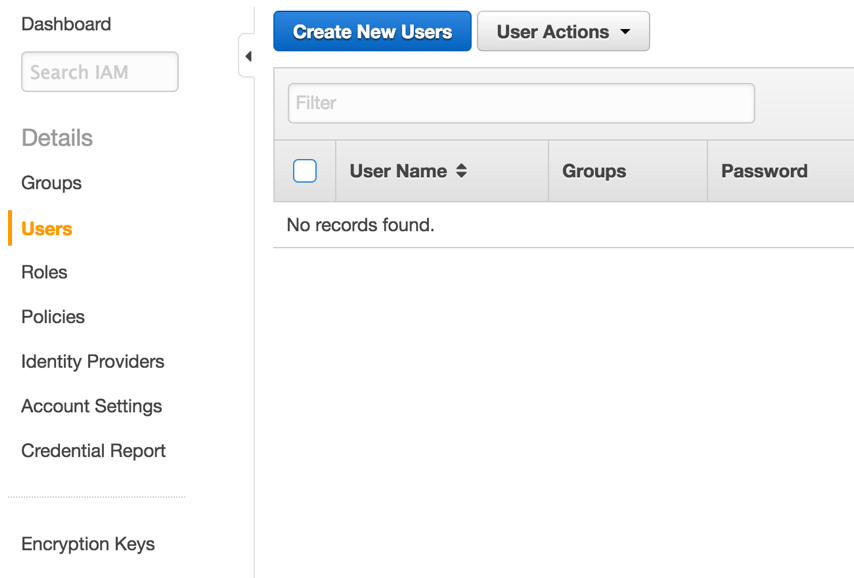
Note that you will need **a credit card associated with your account to download the data, although it is almost free ($0.01 /GB)**.

# Set up a new user in your AWS account

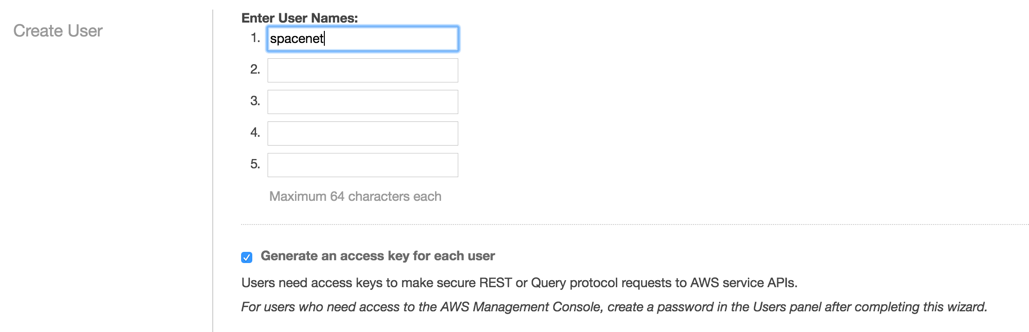
**Step 1.** Once you have an AWS account, you can sign in to your console. Your console looks like the following figure. Click on the “**Identity & Access Management**” as highlighted in the figure.



**Step 2.** Then, on the left side of the screen, you will find a list of details as follows. Please click on the “**Users**”. And then click on the “**Create New Users**” on the right.



**Step 3.** Then, enter **a new user name**, for example “spacenet”, with **“Generate an access key for each user” option selected** (it is selected in default), as shown in the following figure. Click the “**create**” button in the bottom-right corner.



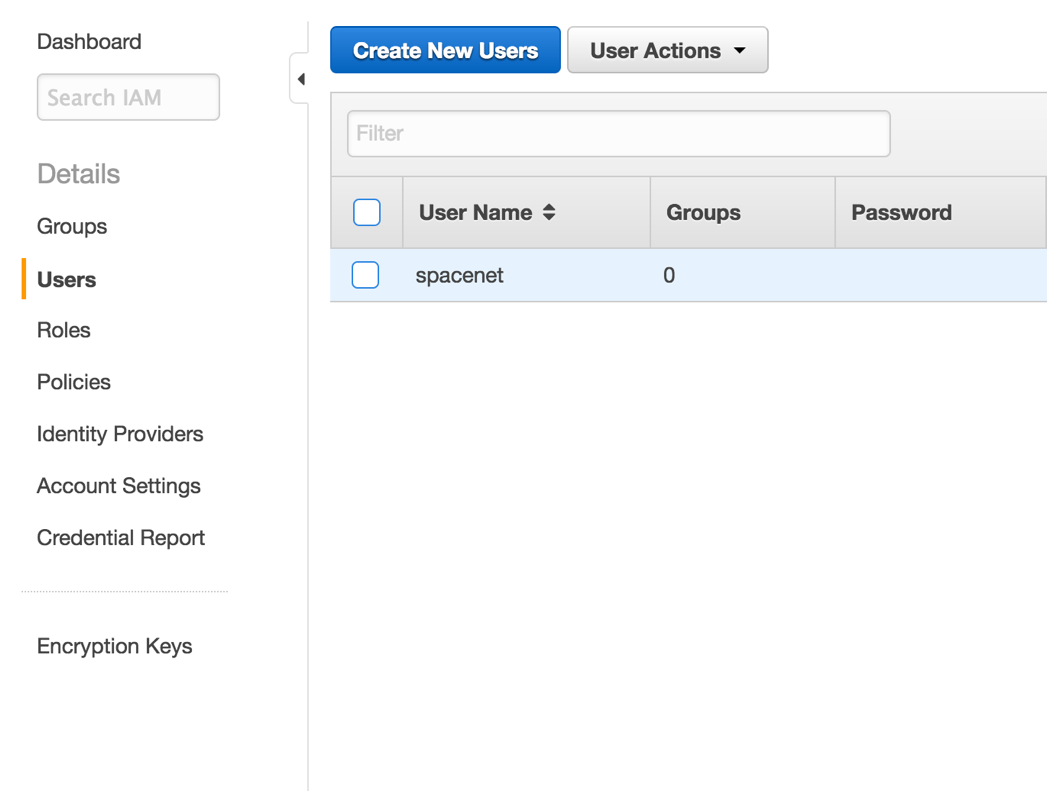
**Step 4.** Please save your credentials file and then close. To save this information, you can either click on “**Show User Security Credentials**” or “**Download Credentials**”. We will use this pair of “**Access Key ID**” and “**Secret Access Key**” later. Because of privacy issue, there is no figure instruction there ☺.

Now, we are going to change the permission of your user page. In our example, it is **IAM > Users > spacenet**. If you use a different username, the last click should be changed accordingly.

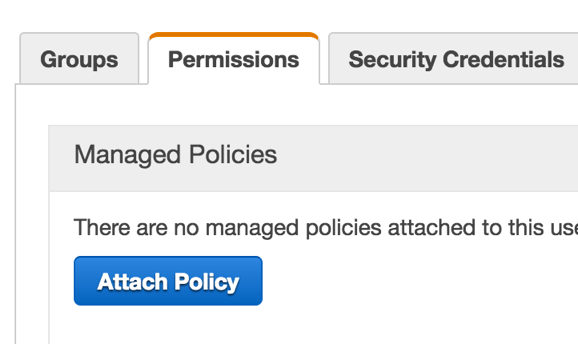
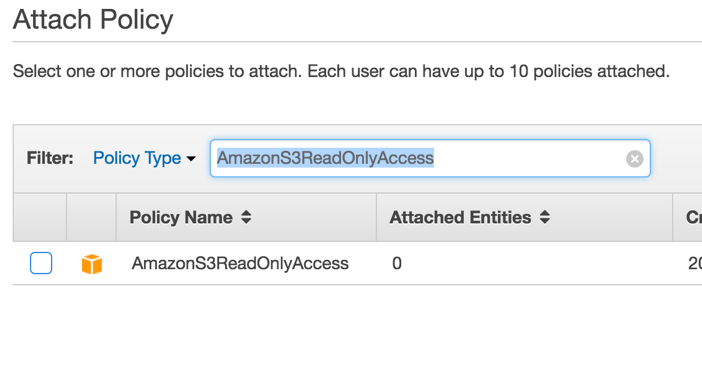
**Step 5.** First, click “**Services**” on the top of the web page, and click “**IAM**” as highlighted in the following figure:



**Step 6.** Click “**Users**” and select the new account (e.g., **spacenet**) you just created.



**Step 7.** Click “**Permissions**”, “**Attach Policy**”, find “**AmazonS3ReadOnlyAccess**”, check the box, and click “**Attach Policy**”.



# 

# Get a Linux or Mac OS

If you don’t have a Linux or mac OS, [Virtual Box](https://www.virtualbox.org/wiki/Downloads) is a free way to get such an OS without purchasing a new machine. And [Ubuntu Desktop](https://www.ubuntu.com/download/desktop) is fairly easy to setup. There are many instructions on how to install Ubuntu in virtual box. One of the instruction could be from [Ask Ubuntu forum](http://askubuntu.com/questions/142549/how-to-install-ubuntu-on-virtualbox).

# Install AWS CLI

## Mac and Linux

First, you need to install “**pip**”.

* On mac, just type “sudo easy\_install pip” in your terminal and press enter.
* In Linux, it depends on your detailed version. Two useful references: [Ubuntu 16.04](https://www.rosehosting.com/blog/how-to-install-pip-on-ubuntu-16-04/) and [Ubuntu 10.10 and older](http://www.saltycrane.com/blog/2010/02/how-install-pip-ubuntu/).

Second, type **“sudo pip install awscli”** in your terminal.

## Windows

Download the AWS CLI installer from [this page](https://aws.amazon.com/cli/) and follow the instructions.

Alternatively you can use Virtual Box to obtain an Ubuntu installation (see above at ‘Get a Linux or Mac OS’) then follow the instructions above for ‘Mac and Linux’.

# Configure AWS CLI

Next, configure aws cli with your credentials saved before. Type “**aws configure**” and then enter your keys:

* AWS Access Key ID [None]: XXXXXXXXXXXXXXXX
* AWS Secret Access Key [None]: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
* Default region name [None]:
* Default output format [None]:

Last, test your connection with the command from [Amazon Web Services S3 repository](https://aws.amazon.com/public-data-sets/spacenet/).

aws s3api get-object --bucket spacenet-dataset --key manifest.txt --request-payer requester manifest.txt

This command will download a file “manifest.txt” in your current working directory.

# Download Single File

If you want to first take a glance at the data by downloading a single file, you may follow the instructions below.

* Open the manifest.txt
* Find your interested file. For example *“./AOI\_1\_Rio/srcData/mosaic\_3band/013022223310.tif*”
* First create the corresponding directory. In our example, the directory is “*./AOI\_1\_Rio/srcData/mosaic\_3band/*”
* Type the following command in terminal to download the file.

**Template:**

aws s3api get-object --bucket spacenet-dataset --key $PATH --request-payer requester $PATH

**Example:**

aws s3api get-object --bucket spacenet-dataset --key AOI\_1\_Rio/srcData/mosaic\_3band/013022223310.tif --request-payer requester AOI\_1\_Rio/srcData/mosaic\_3band/013022223310.tif

# Download Batch Data

We’ve prepared a python 2.7 script for you to download the data. Simply type “**python download-dataset.py**” in your terminal. Make sure the “manifest.txt” file is in the same directory. If you want to download only a part of data, please remove those uninteresting files in “manifest.txt” while keep all directory names there.

Note that

* The wait time might be long.
* You get charged **$0.09 for every GB** of data you download for this much data, but it's still nearly free.

# Minimal data set

The spacenet-dataset bucket contains more data than what is absolutely necessary to participate in this contest. While you may find useful to obtain all data that’s available it is enough to download the following files:

|  |  |
| --- | --- |
| ./competition1/spacenet\_TrainData/3band.tar.gz | Contains 6940 3-band image chips as training data. Size: ~2.5GB |
| ./competition1/spacenet\_TrainData/8band.tar.gz | Contains 6940 8-band image chips as training data. Size: ~890MB |
| ./competition1/spacenet\_TrainData/vectordata/summarydata.tar.gz | Contains the ground truth data corresponding to the training data. You’ll need the AOI\_1\_RIO\_polygons\_solution\_3band.csv file from this package. Size: ~78MB |
| ./competition1/spacenet\_TestData/3band.tar.gz | Contains 2795 3-band image chips as testing data. Size: ~1GB |
| ./competition1/spacenet\_TestData/8band.tar.gz | Contains 2795 8-band image chips as testing data. Size: ~370MB |

The ./competition1/spacenet\_TrainData folder contains the training data that is available for this contest: 3-band and 8-band images and the corresponding ground truth building footprints.

The ./competition1/spacenet\_TestData folder contains the testing data as 3-band and 8-band images. For this data set ground truth is not provided, your algorithm should generate the corresponding building footprints. In the provisional testing phase you should create a CSV file that contains the polygons representing these footprints. See the problem specification for the exact format of this file and for details on how you should submit it for testing.