Data Lakes Project

Helen Chung | OPLC
Data Science Fellow ’21
Demo Day
August 4th, 2021

Objective

- We want to look into the automation of acquiring accurate and precise pricing data off of websites through the Data Lakes project.
- In our case, we wanted to build a model that would scrape a calendar-based website.

Background

- Currently, getting price data off of calendar-based websites are done manually.
- This proof of concept showcases one example of many automation transitions that could happen in the future and the capabilities of AWS in a government/BLS-specific setting

Tools Used

- Selenium (Python package)
- SageMaker, S3, Lambda (AWS Tools)

Selenium

Selenium is a package that is used to automate web browser interaction from Python. It requires a WebDriver to interact with a browser (e.g. Chrome, Safari, Firefox, Microsoft Edge). It is often used in QA testing.

Implementation: Selenium

	capture_date	ticket_type	num_days	park_type	date_of_entry	price	tax	availability	currency
0	2021-08-02	adult	1		2021-08-02	null	null	False	null
1	2021-08-02	adult	1		2021-08-03	$124	null	True	USD
2	2021-08-02	adult	1		2021-08-04	$124	null	True	USD
3	2021-08-02	adult	1		2021-08-05	$139	null	True	USD
4	2021-08-02	adult	1		2021-08-06	$139	null	True	USD

Capture_date: When the data was gathered

Ticket_type: Adult or Child tickets

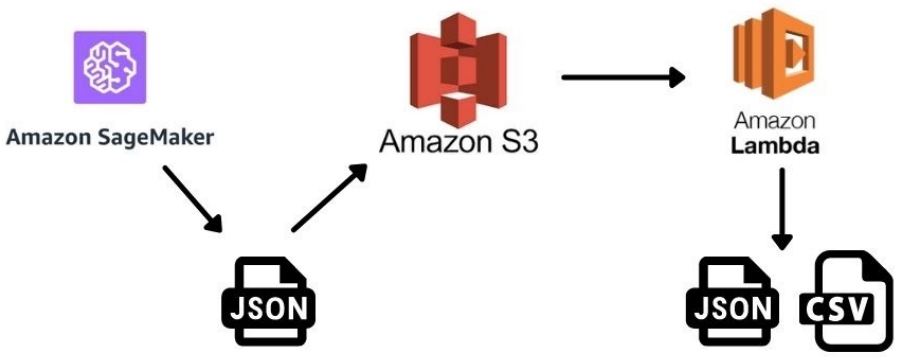
Date_of_entry: Date corresponding to the price

AWS

- **SageMaker**: A group of services on AWS that helps process and clean data.
- **Lambda**: Serverless computing platform that runs after a specified event.
- **S3 (Simple Storage Service)**: Easy and accessible way to store and manage data

AWS

- **SageMaker**: A group of services on AWS that helps process and clean data.
- **Lambda**: Serverless computing platform that runs after a specified event.
- **S3 (Simple Storage Service)**: Easy and accessible way to store and manage data



Implementation: Lambda (1)

capture_date	ticket_type	num_days	park_type	date_of_entry	price	tax	availability	currency
2021-07-27	adult	1		2021-07-30	$154	NaN	True	USD
2021-07-27	adult	1		2021-07-31	$154	NaN	True	USD
2021-07-27	adult	1		2021-08-01	$154	NaN	True	USD
2021-08-02	adult	1		2021-08-02	$139	NaN	True	USD
2021-08-02	adult	1		2021-08-03	$124	NaN	True	USD
2021-08-02	adult	1		2021-08-04	$124	NaN	True	USD
2021-07-27	adult	1		2021-08-05	$139	NaN	True	USD
2021-08-02	adult	1		2021-08-06	$139	NaN	True	USD

We want a file with the most recent price listings. Prices may change over time, or it may appear after being marked as “sold out” (or vice versa).

Implementation: Lambda (2)

ticket_type	num_days	park_type	date_of_entry	average_price	currency
adult	1		6/2021	131.50	USD
adult	1		6/2021	179.00	USD
child	1		6/2021	129.00	USD
child	1		6/2021	0.00	USD
adult	1		7/2021	142.39	USD
adult	1		7/2021	197.39	USD
child	1		7/2021	142.39	USD
child	1		7/2021	197.39	USD
adult	1		8/2021	132.23	USD
adult	1		8/2021	187.23	USD
child	1		8/2021	132.23	USD
child	1		8/2021	187.23	USD

We want a file of the monthly average prices for each specific item.

Next Steps

- Explore other tools in AWS that can improve automation or other similar processes
- Expansion to other websites

Contact Information

Helen Chung | OPLC
Data Science Fellow ‘21
Demo Day
Chung.helen@bls.gov