Konstantino Sparakis : U47131572
Shuwen Sun: U74918696
November 4, 2016
CS591

Mid-progress report

GITHUB: https://github.com/AWS-Spot-Analysis/spot-analysis

Attached to the email are:
Sample data from api,
Sample data from archives,
And our working code in jupyter notebook

Abstract

We are analysing the price marketplace of AWS EC2 spot instances. Through our research, we plan to study the marketplace which consists of bidding for virtual resources. Our goal, is to discover patterns and test hypothesis using the data provided from this market place. We hope to be able to solve question like are we able to predict the volatility of the bid market for certain machines given. We hope that our analysis will help people understand the marketplace better allowing people to save money, and give people a starting point regarding future research.

Introduction

Our dataset is a time series of bid pricing for Amazon clouds EC2 Spot instances. EC2 is a service that allows you to purchase and use a custom virtual machine on an hourly basis that runs in a remote server and location, the cloud. EC2 virtual machines consist of different specs, allowing users to choose everything from ram, cpu's, ssd memory, OS and more options. Amazon also has multiple data centers throughout the world and these are considered regions.

What amazon introduced with Spot instances is that since not all their computing resources are being used up at any given moment. You can bid on these leftover machine resources starting at about 1/10th the original price. As the resources available is variable with decrease in supply, the demand goes up and so does the price of the resources. If you get outbid your instance is shutdown after 2 minutes.

Data Collection

What our approach has consisted of is first off hunting for data, aws only provides 90 days of most recent data, archives exist but finding ones that were still serving their data was difficult but we managed to pull through. We obtained data ranging from 2014-04-17 to 2015-08-24 that looks like such:

Title	Price	Timestamp	Machine type	OS	AWS Region
SPOT INSTANCE PRICE	0.107100	2014-03-27T07:33:04-0700	c3.2xlarge	Linux/UNIX	us-west-1a

We also have used the aws terminal client and api to grab the previous 90 days. Which a sample of this is included in our report. It comes in JSON format so some more processing needs to be done but this is rather simple.

Using this command:

```
aws ec2 describe-spot-price-history --instance-types m1.xlarge --start-time
2014-01-06T07:08:09 --end-time 2014-01-06T08:09:10
```

Retrieving:

```
...
{
        "Timestamp": "2016-11-06T08:01:04.000Z",
        "ProductDescription": "SUSE Linux (Amazon VPC)",
        "InstanceType": "m1.xlarge",
        "SpotPrice": "0.133500",
        "AvailabilityZone": "us-east-1d"
 }
...
```

Hypothesis

Now with this data is where it gets tricky and creative. We have decided to only look at machines whose operating system is Linux, this is due to more data being available and to tighten our scope. We have to in a sense massage the data in order to start seeing what techniques return interesting results worth reporting. In order to give ourselves a direction we have come up with three hypothesis we want to answer:

Hypothesis #1
Can we predict future price of a spot instance given previous history and how other vm's are reacting?

To achieve the goal of prediction, we are expecting to do pattern matching from the collected dataset. In this case, whenever users make a bid, we can based on the resources types, time or day, and the trending price to do pattern matching. We will be able to provide a prediction if we can shoot a pattern.

Expecting tools are supervised and unsupervised learning algorithms, e.g. classification and clustering methods.

Hypothesis #2
For each machine type there exists a region that is more favorable to use, as the market volatility is very low and the prices tend to stay cheaper than the other regions.

With in proving this hypothesis users will be able to find the best region they should be bidding in, as long as latency is not an issue for them.

Data Science tools & Techniques:
We can use clustering and classification methods.

Hypothesis #3
There exists some kind of relation between what kind of virtual machines are turning into hotspots.

Say that we establish a line as half price of EC2 instances, it makes sense to pay half price to gain usage of resources but probably not more than ¾. By extracting patterns from the price history, we can study that whether or not there was the case that some resources were becoming hotspot in the spot instances market.
Data Science tools & Techniques:
Time Series, Linear Regression

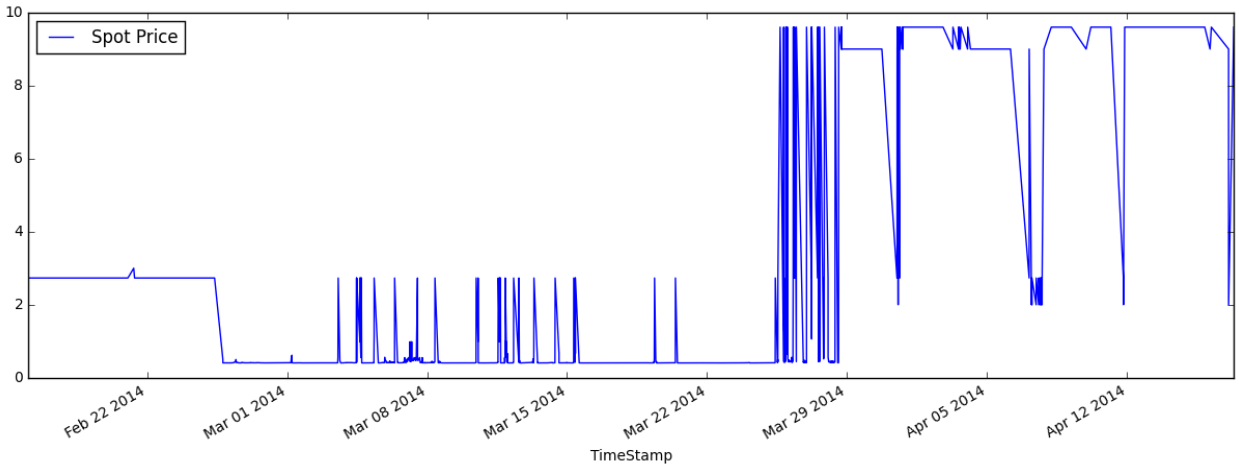
Our goal with our analysis and answering these hypothesis is to help inform people on the environment of the market. We hope to uncover things that will allow people to choose a machine, in an area with a better idea of how likely they are to get outbidded, and what days might be the better days for them to use spot instances to get a cheaper pricing, or is it in fact even worth just consistently paying for spot instance and move away from original instances?

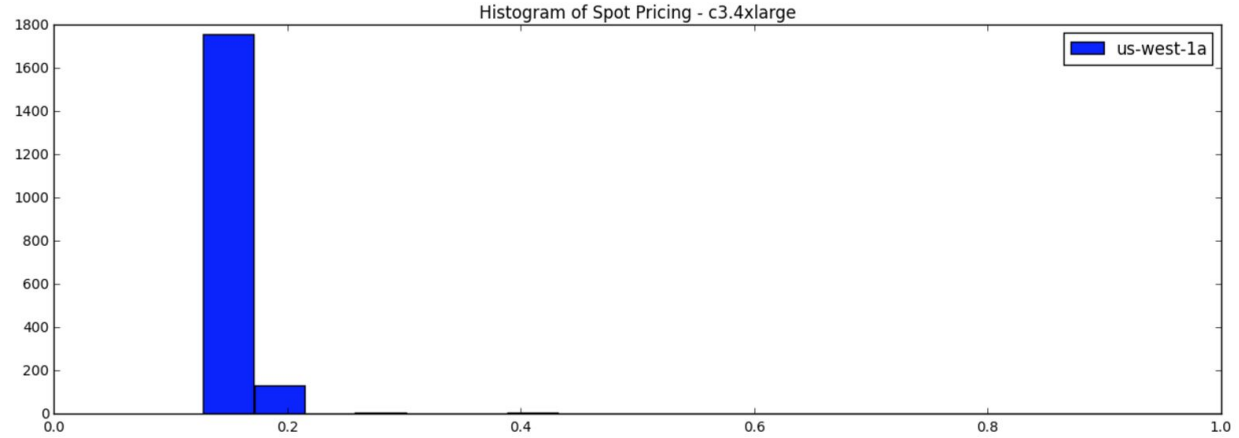
Preliminary Experiments

Some ideas that we have come across, are using measures such as RSI which tells us market momentum and along with this other typical stock market analysis tools to help us get a clearer picture of our data and see if we can learn anything from using these tools on aws's market. But as an initial analysis we have created code that parses the 2014-04-17 to 2015-08-24 dataset. And we started graphing it in order to get a more visual explanation of our data and get a better idea of things that we could be looking for.

In our mid-progress work, we take the virtual machines in region "us-west-1a" of type "c3.8xlarge". Obviously, the marketplace may various and the type of spot resources may also be a huge factor to the conclusion. We stick to one type of machine and region so that we can establish some pattern and get some useful knowledge before jumping into the permutation of changing factors.

Here are some example graphs:





Results and Discussion

To be completed with the final report.

Conclusion

To be completed with the final report

Below is our in-progress work:
AWS-Spot-Analysis/spot-analysis