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

Project proposal

The Dataset & It's Nature

The datasets we are used to study our topics are Amazon EC2 spot instance prices. Nature of the datasets are actually the bidding people made succeed, they also revealed some trends about what type and where those virtual machines were needed. Datasets contain information about different prices at different time, AWS regions, and various machine types. We have apprehended the recent 90 days' data from Amazon, and this dataset should be our main focus. We have also obtained a full dataset from 2014-04-17 to 2015-08-24.

Below is the example format

Title	Price	Timestamp	Machine type	OS	AWS Region
SPOT INSTANCE PRICE	0.107100	2014-03-27T 07:33:04-070 0	c3.2xlarge	Linux/UNIX	us-west-1a

Expected Analysis

Our expected analysis will work as follows. We will start off with some ideas and slowly expand to encompass more as time permits. Our analysis will work by us asking questions of interest and seeing if we can answer them, with the data we have and tools and techniques we know.

Operating System Analysis
We will only be studying Linux/UNIX, in the AWS environment because we were able to acquire more data on this, and also it is the most widely used operating system in AWS so it makes sense to make this the center of our attention.

Our Hypothesis:

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.

=Potential data science method for this one includes: Time Series, Linear Regression

Applications of Our Research

Our dataset is a time series of bid pricing for Amazon clouds EC2 Spot instances. EC2 is a cloud virtual machine that you can purchase on an hourly basis. What amazon introduces with Spot instances is that since not all their computing resources are being used up at any given moment. You can bid on machine resources 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.

Our analysis will help people be more informed 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.

Expected Results

We have already found one year's worth of data from 2014-15, but will also be scraping the website to get the most recent data, but in fact scraping won't be necessary because you can obtain this data just using aws's api to request it.

Depending on the analysis we might be forced to acquire more data. We also expect to have to get regular AWS ec2 pricing data as we might have to adjust for deflation as the prices have been dropping over time.

We hope the data is sufficient that the analysis will yield promising results. We expect to find answers to hypothesis and yield results that will assist people make more educated bids in location and machine types.